Engineering 25 (2023) 110–119

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Article

1000× Faster Camera and Machine Vision with Ordinary Devices

Tiejun Huang, Yajing Zheng, Zhaofei Yu *, Rui Chen, Yuan Li, Ruiqin Xiong, Lei Ma, Junwei Zhao, Siwei Dong, Lin Zhu, Jianing Li, Shanshan Jia, Yihua Fu, Boxin Shi, Si Wu, Yonghong Tian

School of Computer Science, National Engineering Research Center of Visual Technology, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Received 19 February 2021
Revised 3 January 2022
Accepted 5 January 2022
Available online 12 April 2022

Keywords:
Vidar camera
Spiking neural networks
Super vision system
Full-time imaging

ABSTRACT

In digital cameras, we find a major limitation: the image and video form inherited from a film camera obstructs it from capturing the rapidly changing photonic world. Here, we present vform, a bit sequence array where each bit represents whether the accumulation of photons has reached a threshold, to record and reconstruct the scene radiance at any moment. By employing only consumer-level complementary metal–oxide semiconductor (CMOS) sensors and integrated circuits, we have developed a spike camera that is 1000× faster than conventional cameras. By treating vform as spike trains in biological vision, we have further developed a spiking neural network (SNN)-based machine vision system that combines the speed of the machine and the mechanism of biological vision, achieving high-speed object detection and tracking 1000× faster than human vision. We demonstrate the utility of the spike camera and the super vision system in an assistant referee and target pointing system. Our study is expected to fundamentally revolutionize the image and video concepts and related industries, including photography, movies, and visual media, and to unseal a new SNN-enabled speed-free machine vision era.

1. Introduction

Is a digital camera truly digital? The typical answer is yes, as imaging on film is replaced by imaging with charge-coupled device (CCD)/complementary metal-oxide-semiconductor (CMOS) sensors and digital circuits. However, the essence of the digital camera remains in the analog age, as it indiscriminately inherits the image and video form, which are necessary to record the temporal dynamics of light on film [1–3] but are not necessary for a purely digital system. In fact, an image cannot record the change in light during the exposure time, and a video even misses all of the dynamic information between two neighboring exposures. Furthermore, the frame rate of cost-effective consumer-level cameras is only tens of hertz, making capture of high-speed scenes impossible. In contrast, high-speed cameras can reach a time sampling frequency of thousands or even tens of thousands of hertz, but they require specialized sensors and shutters that are highly expensive [4,5]. Therefore, the image and video form has become the greatest obstacle for digital cameras to capture the fast-changing photonic world.

In this study, we propose a revolutionary visual representation, called vform, that breaks the conventional frame-based representation, allowing cost-effective high-speed cameras to be made. Inspired by the sampling mechanism of primate fovea [6,7], vform takes advantage of spike sequences to represent the changes in light in the spatial–temporal domain and can accurately retain the timing of physical optical flow. This brings the ability to reconstruct the scene radiance at any given moment, which is called full-time imaging.

Based on the new visual representation model, we develop the VidarOne chip and spike camera with the same CMOS sensors and consumer-grade integrated circuits as in traditional cameras [8]. A spike is generated when the accumulated intensity collected by the photosensitive devices exceeds a given threshold. The photoelectric conversion speed of these photosensitive devices is approximately 10 ns, which is six orders of magnitude faster than that of a human retina [9]. Therefore, while having similar mechanisms, the spike camera avoids the speed limitation of biological vision. The first spike camera we developed has a time sampling frequency of 40 000 Hz, which can be used to implement high-speed imaging 1000 times faster than that of human vision and conventional cameras.

* Corresponding author.
E-mail address: yuzf12@pku.edu.cn (Z. Yu).

https://doi.org/10.1016/j.eng.2022.01.012

The spike streams generated by the spike camera have a clear physical meaning; that is, they encode the spatial–temporal visual information of the input scene and can thus be used to perform high-speed vision tasks. However, conventional machine vision methods based on artificial neural networks (ANNs) [10] cannot process these spike streams in real time because they must first convert the spike streams to images (40 000 frames per second) and then process them frame by frame. In contrast, we find that spiking neural networks (SNNs) [11,12] can naturally process the output spike streams of a spike camera in real time. Using this approach, we developed an SNN-based supervision system that combines the speed of the machine and the mechanism of biological vision [13–18]. The vision process can be understood as the flow of spike sequences within the SNN; thus, the processing speed only depends on the physical properties of the SNN. We realized real-time processing of 40 000 Hz vform spike streams in the supervision system with ordinary central processing units (CPUs) and achieved high-speed moving object detection and tracking that is 1000 times faster than that of human vision. In the future, by using SNN hardware and higher speed spike cameras, we can implement object detection, tracking, prediction, and recognition at electrical speeds and achieve superhuman vision faster by more orders of magnitude. All we need are regular consumer-grade optoelectronic devices and circuit technologies that are widely used today.

2. Methods

(1) **Visual texture reconstruction.** The texture from window (TFW) method obtains the pixel value (proportional to the scene radiance) by calculating the number of spikes in a time window. Specifically, a moving time window collects spikes in a specific period. By counting these spikes, the pixel value is estimated by:

$$P_{t_i} = \frac{N_w}{w} \cdot C \tag{1}$$

where P_{t_i} refers to the pixel value at moment t_i; w is the size of the time window that contains the previous w moments before t_i; N_w is the total number of spikes collected in the time window; and C refers to the maximum dynamic range of the reconstruction. The texture from interspike interval (TFI) method assumes that the scene radiance $\bar{I}$ is a constant in a short period. According to the spike camera mechanism, the spike generation condition can be simplified as $\bar{I}\Delta t \geq \phi$, where Δt is the interspike interval obtained by calculating the time between two neighboring spikes and ϕ denotes the trigger threshold. Thus, the pixel value can be estimated with two spikes (i.e., one interspike interval):

$$P_{t_i} = \frac{C}{\Delta t_i} \tag{2}$$

where Δt_i represents the interspike interval corresponding to moment t_i.

We test the proposed image reconstruction algorithms and compare it with conventional camera. We build a hybrid camera system consisting of the spike camera, conventional camera, and a beam splitter. Two cameras can record the same scene through the beam splitter. We employ two no-reference image quality assessment metrics, namely two-dimensional (2D) entropy and standard deviation (STD). 2D entropy uses both the gray value of a pixel and its local average gray value to evaluate the amount of information carried by the image, larger 2D entropy means more information. STD evaluates the contrast of the image, and larger STD means higher contrast. As shown in Table 1, our reconstruction methods achieve better results than conventional camera in all two metrics.

(2) **Dynamic connection gate.** The dynamic connection gate is based on short-term plasticity (STP), which refers to the short-term change in synaptic strength (usually between tens to thousands of milliseconds), also known as the dynamic connection between neurons [19,20]. When a postsynaptic neuron receives a sequence of action potentials from a presynaptic neuron, the postsynaptic potential (PSP) changes according to:

$$\mathrm{PSP}(t) = A \cdot x(t) \cdot u(t) \tag{3}$$

where A is the maximum current value that an action potential can trigger on a postsynaptic neuron; $x(t)$ $(0 < x(t) < 1)$ represents the remaining number of available neurotransmitters in the axon terminal at time t; and $u(t)$ denotes the release probability of neurotransmitters in the axon at time t. When a postsynaptic neuron receives a sequence of action potentials with fixed frequency from a presynaptic neuron, the PSP converges to a stable state after several spikes arrive [21] (Fig. 1(a)). If the spike frequency changes, then the PSP will fluctuate around a stable value (Fig. 1(b)). By taking advantage of the sensitivity of the STP to the release time mode of the input spike streams, the spike streams generated by the background or static areas can be filtered, and only the spike streams generated by the moving object are retained.

(3) **Detection and tracking.** The neuron in the filter layer is connected to nine adjacent leaky integrate-and-fire (LIF) neurons in the detection layer. Each LIF neuron accumulates current from presynaptic neurons and fires when the membrane potential reaches the threshold. As only the area corresponding to a moving object in the detection layer can generate spike streams, each moving object can be found by detecting the connected area of the firing neurons. The tracking-by-detection method is utilized to track different moving objects. To evaluate algorithm accuracy, we use the detection success rate (DSR) to measure the effect of object detection, and multiple-object tracking accuracies (MOTA), false positive (FP), miss detected (FN), identifier switches (IDS) to evaluate the effect of object tracking. The results are shown in Table 2, we can find that our algorithm can achieve good performance with low power.

(4) **Continuous attractor neural network (CANN) for prediction.** A CANN is a canonical network model for neural information representation. A previous study has revealed that by adding negative feedback to neuronal dynamics, a CANN can track a moving object anticipatively with an approximately constant leading time [22]. Based on this, we develop a CANN model for anticipatively tracking a fast-moving object in real-world applications, with the visual inputs coming directly from the spike camera.

(5) **Object recognition.** The synaptic weights of the recognition SNN are trained with the backpropagation (BP)-spiking-timing-dependent plasticity (STDP) learning rule, which is derived from Tavanaei and Maidi [23]. Here, we use multiple spike neurons to

Table 1
Comparison among TFI, TFW, and conventional camera.

Index	Scene	TFI	TFW	Conventional camera
STD	Motion	73.81	74.33	73.25
	Static	73.82	74.29	73.44
2D entropy	Motion	12.86	13.17	11.54
	Static	12.83	13.21	11.78

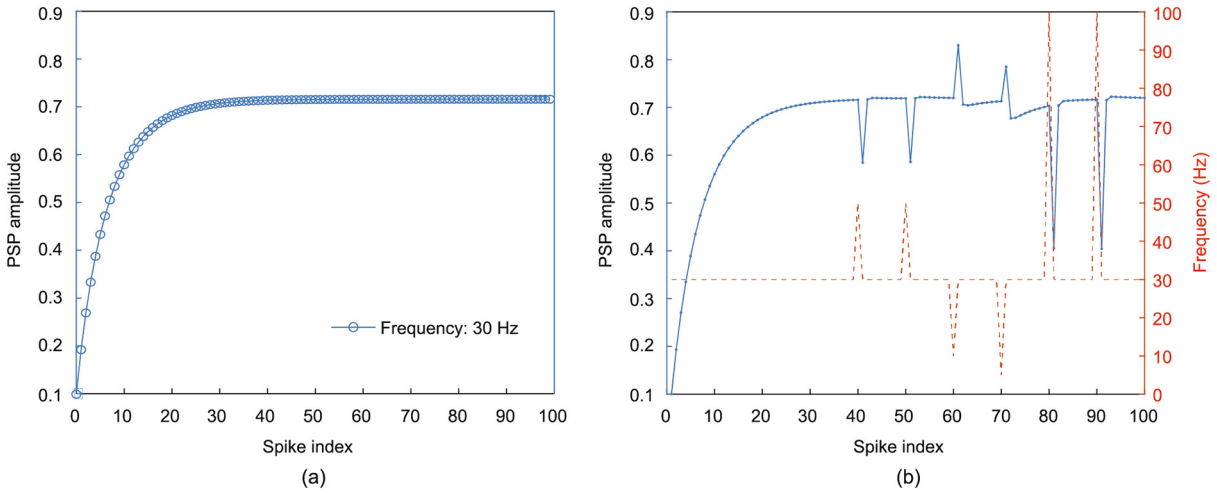

Fig. 1. PSP changes with respect to the spikes from a presynaptic neuron. (a) The PSP amplitude converges to a stable state for the input spikes with a fixed frequency of 30 Hz; (b) the PSP amplitude (blue curve) fluctuates around a stable value for the input spikes with varying frequency (red curve).

Table 2
Accuracy of detection and tracking. Quantitative evaluation of detection and tracking.

DSP	MOTA	FP	FN	IDS	Speed (Hz)	Power (W)
100%	96.32%	23	23	0	20 811	2.254

represent one category in the last layer of recognition SNN. Specifically, m neurons are divided into n groups to represent n categories ($m = kn$, where k denotes the number of neurons per category). The real spike data generated from the spike camera mixed with the simulated data generated from Spike-Sim are used as the training dataset. In the training process, the classification neuron with the maximum membrane potential in the target group updates its synaptic weights according to STDP [24], while the misfired classification neurons with the maximum membrane potential in the nontarget groups undergo anti-STDP [25]. Then, the modulation of synaptic weights is backpropagated layer by layer according to presynaptic activities.

(6) **Estimation of velocity.** As the angular velocity of the fan is 2400 revolutions per minute (rpm) and the distance from the center of the characters to the center of the fan is 0.12 m, the linear velocity of the characters on the fan can be estimated as $2400/60 \times 2 \times \pi \times 0.12 \approx 30$ m·s^{-1}. Considering that the distance between the fan and the spike camera is 0.75 m, the vidar camera and the super vision system can detect, track, and recognize a moving object with a linear velocity of 40 m·s^{-1} within 1 m in real time according to the central perspective principle.

(7) **Spike camera high-speed spike dataset (VHSSD).** This dataset includes ① spike streams of high-speed moving targets captured with a static vidar camera (Class A) and ② spike streams of natural scenes captured with a high-speed moving spike camera (Class B). Class A contains a moving car, a rotating disc, a rotating fan, and a bursting balloon, while Class B contains train, forest, viaduct bridge, and railway scenes (more details can be found in Table 3). We also provide SpikePlayer for playback of the spike sequences.

(8) **SpikePlayer.** This visualization software can play real and simulated spatial–temporal spike streams (i.e. dat files) recorded by the spike camera, providing high frame rate videos reconstructed with the proposed TFW and TFI. SpikePlayer supports various resolutions, such as 400 × 250, and even extends the simulator's compatibility.

Table 3
Unified description of the SCHSSD.

Sequence	Length (s)	Spike number
Class A: moving target		
Moving car (100 km·h^{-1})	0.20	102206031
Rotating disc (7200 rpm)	3.84	535852602
Rotating fan (2400 rpm)	2.00	407620564
Bursting balloon	0.10	6351184
Class B: moving spike camera		
Moving train (350 km·h^{-1})	0.20	42898223
Forest	0.22	93319068
Viaduct bridge	0.22	136859111
Railway	0.22	87866720

(9) **Spike-Sim.** Spike-Sim is a simulator of the spike camera used to simulate arbitrary camera motion and object motion in three-dimensional (3D) scenes and provides reference images and additional information, including camera pose, object velocity, and so forth. This simulator integrates the principle of spike camera theory and multiple rendering engines, including a fast and custom renderer developed based on Open Graphics Library (OpenGL) that can render and generate spike streams in real time and a photorealistic render based on Blender's Cycles engine.

3. Results

3.1. Vform: A new and more natural visual form

Before we introduce the new visual representation called vidar, we briefly review the concepts of images and videos. For a large number of photons traveling within a camera's viewing frustum (Fig. 2(a)), the camera will acquire images at time t_1, t_2, t_3, …, t_n (the time interval is $1/f$ seconds) according to the predetermined frame rate f. During image capturing, all the photosensitive units simultaneously capture photons over a duration of Δt (known as

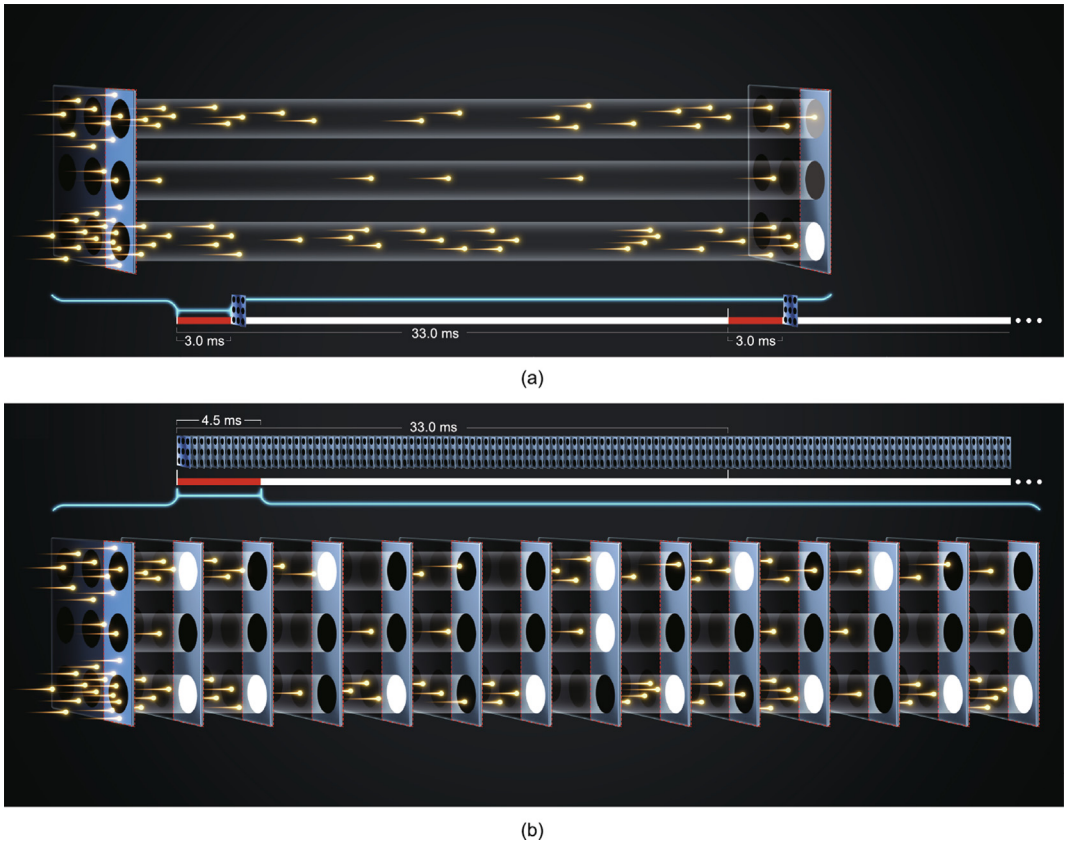

Fig. 2. Overall comparison of image and vform in terms of visual information representation. (a) Visual representation by images and video. The photosensitive units (three circles) capture a group of photons (yellow shooting stars) and output the accumulated intensity (shown by the brightness of the circle) during the exposure time of 3.0 ms (red line). An image corresponds to the intensity distribution according to the spatial arrangement of the photoreceptors, and a video is generated by acquiring images every 33.0 ms according to the predetermined frame rate f = 30 Hz. Note that the visual information during the 30.0 ms interval (white line) is completely lost. (b) Visual representation by vform. The photosensitive units (three circles) continuously capture photons and generate a spike (white circles) when the accumulated intensity exceeds a given threshold (here, the threshold is four photons). Vform is a bit sequence array arranged in accordance with the spatial arrangement of the device, where a bit 1 indicates that a spike is generated by the photosensitive unit at that moment, and a bit 0 indicates that the unit is in the cumulative state. The vform for the three photosensitive units here is $\begin{bmatrix} 01010001010100 \\ 00000001000000 \\ 01101010101011 \end{bmatrix}$.

the exposure time, $\Delta t < 1/f$) and then record the accumulated intensities. The distribution of intensities according to the spatial arrangement of the photoreceptors forms the image, and a sequence of images arranged at equal time intervals is called a video. From the perspective of the plenoptic function [26], what is recorded in the image is not the moment at time t_1, t_2, t_3, ..., t_n, but the accumulation of physical processes that last Δt. For a video, the cumulative time (exposure time) Δt used to acquire each frame is less than or equal to the time interval $1/f$ between two frames of the video, which means that the information in period $\frac{1}{f} - \Delta t$ is completely lost, and the motion process during time interval Δt is also "squashed" into the image and lost. Thus, the temporal domain sampling of the video is not a complete sampling of the physical process.

The synchronous exposures and the same exposure time in images and videos impede the ability of digital cameras to capture the rapidly changing photonic world. However, such a design is not necessary. Here, we introduce vform, a new visual representation that can better capture the temporal domain changes in light by utilizing a new temporal domain sampling mechanism and allowing asynchronous exposures. Here, vform, as the combination of "vi-" (visual) and "form", without any capitalization, is coined to define a new form of visual information to replace video.

Specifically, vform is no different from traditional images and videos in terms of spatial sampling. Vform uses the same photosensitive devices (i.e., well-known CMOS or CCD sensors, as in traditional cameras). Therefore, vform also takes advantage of the spatial arrangement of a lattice to represent spatial information (Fig. 2(b)). The fundamental difference between vform and video is the use of a new temporal domain sampling method. All the photosensitive units continuously capture photons instead of being synchronously exposed with the same exposure time. When the accumulated intensity exceeds a given threshold, a spike is generated (Fig. 2(b)). The spike and the duration required to generate this spike are called a vit. The vits generated by each photosensitive unit are arranged in sequence according to chronological order. The simplest representation of vits is a bit stream, where 1 indicates that a spike appears at that moment, and 0 indicates that the unit is in the cumulative state. A bit 1 and all the 0 s between this 1 and the previous 1 constitute a digital vit. Each photosensitive unit can generate a spike stream, and the spike streams generated by all the photosensitive units are arranged in accordance with the spatial arrangement of the device to form a spike stream array, that is, vform.

The outstanding advantage of vidar over video is that the temporal domain change in light at each sampling position is effectively retained. With a device that is sensitive to a single photon, a photon can excite a spike. In this case, vform records the exact

and complete physical process. Ordinary photosensitive devices excite a spike only when they capture a set of photons, which is a rough representation of the physical process. However, the time relationship of the physical process is still preserved to the greatest extent, in contrast to when the time relationship is arranged uniformly at tens of hertz through artificial rules such as in video. In fact, the time sensitivity of CMOS photosensitive devices widely used today has reached tens of nanoseconds. With this new model vform, high-speed temporal domain sampling of ten million hertz can be achieved, and extremely fast physical processes can be recorded. Of course, daily vision applications do not require such a high sampling frequency. The first chip we developed set a sampling frequency of 40 000 Hz, which is 1000 times faster than the sampling frequency of human vision and traditional cameras. Clearly, this chip can be utilized to shoot a high-speed rail with a speed of 350 km·h^{-1} and a hard drive rotating at a speed of 7200 rpm.

Vform records fine changes in light at various positions within a certain spatial range, and its physical meaning is very clear. Therefore, it is expected that vform can be used to generate traditional images and videos. In fact, for any given moment, the scene radiance at each position and the pixel value of each pixel can be estimated from the vit covering that moment, and more detailed scene radiance and pixel value can be estimated by referring to the previous and spatially adjacent vits, thereby obtaining fine images at arbitrary moments. The ability of vform to reconstruct the scene radiance at any moment is called full-time imaging or continuous imaging.

3.2. VidarOne chip and spike camera system

The VidarOne chip is developed based on the new visual representation model spike and adopts an asynchronous pixel trigger architecture. As shown in Fig. 3(a), the 400 × 250 pixel array converts the input photons into a spike stream array and utilizes a rolling shutter to detect the responses of all the pixels. After that, the row scanner scans the pixel array row by row. When one row of pixels is selected through the logic control signal, the data are transferred into the digital buffer for parallel readout. To support the high-speed output of the spike stream, the VidarOne chip provides an eight-channel specialized communication interface with a bandwidth of 500 Mb·s^{-1}. The synchronous readout interface is clocked at 20 MHz.

The basic circuit in a pixel, as shown in Fig. 3(b), consists of a spike trigger circuit, a reset circuit, and a readout circuit. The photodiode in the pixel continuously captures photons and converts the incident light illumination into a continuous photocurrent I_{ph}. Thereafter, the photodiode voltage V_{pix} decreases during the collection of photoelectrons. When the photodiode voltage V_{pix} reaches a certain threshold V_{ref}, the output of the comparator toggles, and a flip (spike) signal is generated (see Fig. 3(c) for illustration). The latch synchronizes the flip signal of the comparator under the enable operation of the clock signal (clk). Once the latch detects the flip signal, the photodiode voltage V_{pix} is reset to a predefined reset voltage. Meanwhile, the spike signal is sent to the recommended standard (RS) flip-flop and saved. The row readout signal R_d controls the sequential scanning and readout of the spike streams, and the row reset signal R_{st} is responsible for clearing the signals in the RS flip-flop. The timing diagram of the spike pixel is shown in Fig. 3(d). Under the control of the clk, the spike signal is fixed at a high level lasting 100 ns. For the spike signal generated at time A (see arrow A), the spike is read out after 50 ns by the row readout signal R_d at a high level, considering that the row reset signal R_{st} is at a low level. For the spike signal generated at time B (see arrow B), the RS flip-flop captures the spike signal after 50 ns, as the RS flip-flop is shielded through the reset signal R_{st} at a high level. The spike is read out when the next readout signal R_d arrives. In a readout cycle, only one spike signal can be processed even if two or more spike signals are triggered. The reason is that the RS flip-flop will not respond to the other spikes when it is latched by a spike signal. As the row scanning time is 100 ns, the time resolution of spike streams generated by the 250-row pixel array is 25 μs.

Fabricated using standard 110 nm 1-poly 3-metal process technology, the VidarOne chip occupies a die area of 9.96 mm × 7.10 mm (Fig. 3(e)). Each square pixel has a size of 20 μm × 20 μm and achieves a 13.75% fill factor on the prototype chip. The large-size pixel detector can guarantee a sufficient photodetector area after placement of the metal grid. Under natural light, the chip can provide a high dynamic range of more than 100 dB without using a dynamic range enhancement technique. The energy consumption of the proposed design is approximately 370 mW. A physical view of the packaged VidarOne chip is shown in Fig. 3(f). The placement and routing are carefully designed to minimize the silicon area for the pixel circuits.

The spike camera system, which is composed of a visual information acquisition module, a high-speed sensing module, and a real-time visual computing module, is developed (Fig. 3(g)). The visual information acquisition module converts the input scene into spike streams, which are passed through the sensing module to undergo high-throughput real-time data processing operations and then are sent to the visual computing module through the peripheral component interconnect express bus.

3.3. Visual texture reconstruction with the spike camera

The spike camera has the ability of full-time imaging. The visual textures at any given moment can be reconstructed according to the characteristics of output spike streams (vits), and the dynamic range and quality of reconstructed textures are very flexible. To reconstruct the captured scene and bridge the gap between vidar data and conventional frame-based vision, we propose two visual texture reconstruction strategies, namely, TFW (Fig. 4(a)) and TFI (Fig. 4(b)). More details are in Section 2.

Specifically, the TFW method takes advantage of the principle that the scene radiance is directly proportional to the spike count (firing rate); thus, one can compute the pixel value (proportional to the scene radiance) by using a moving time window to collect the spikes in a specific period (Fig. 4(a)). The reconstruction results are illustrated in Fig. 4(c), where we present a novel spike dataset called SCHSSD (see Section 2). The first row of Fig. 4(c) presents the raw data of the spike camera for eight different scenes and the second row shows the texture reconstruction with TFW. The TFW method is suitable for stationary scenes. In the case of high-speed moving scenes, the scene radiance received by the spike camera changes rapidly. At this time, the firing rate over a period cannot capture this rapid change in the scene radiance, causing blurry imaging (Fig. 4(c) second row). The TFI method is proposed to solve this problem by utilizing the fact that the scene radiance is inversely proportional to the interspike interval (Fig. 4(b)). Thus, only two spikes (i.e., one interspike interval), are needed to estimate the scene radiance in this period, which can match the rapid change in the scene radiance for high-speed moving scenes. In fact, the texture reconstructed with TFI updates the motion nearly synchronously. TFI achieves better results than TFW for high-speed moving scenes (Fig. 4(c) third row). We also compare our construction results quantitatively with that of conventional cameras. As illustrated in Table 1, our reconstruction methods achieve better results than conventional camera.

To facilitate the demonstration of the new idea, we develop a spike camera simulator, Spike-Sim, that can simulate arbitrary camera motion and object motion in 3D scenes and generate reliable

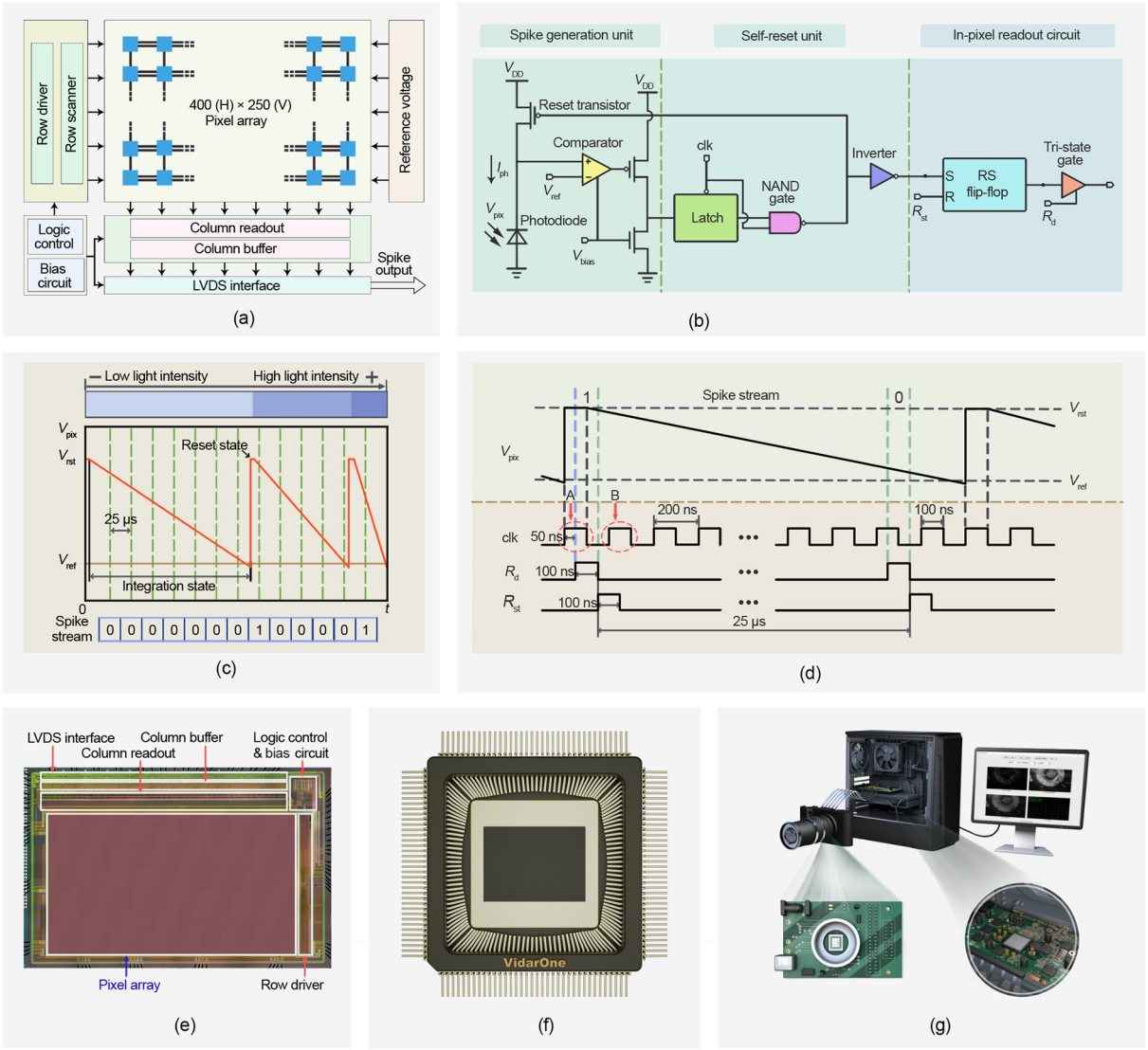

Fig. 3. Design of the VidarOne chip and the spike camera system. (a) Schematic of the chip architecture. It mainly consists of a pixel array, a row scanner with the configured driver, a column readout circuit with an addressable digital buffer, bias/reference circuits, and digital logic control. H: horizontal; V: vertical; LVDS: low voltage differential signaling. (b) The spike generation unit includes three parts: a spike trigger circuit consisting of a photodiode, a reset transistor, and a comparator; a self-reset unit consisting of a latch, a not and (NAND) gate, and an inverter; and an in-pixel readout circuit consisting of a recommended standard (RS) flip-flop and a tri-state gate. V_{DD}: supply voltage; V_{bias}: bias voltage; V_{pix}: photodiode voltage; V_{ref}: reference (threshold) voltage; V_{rst}: reset state voltage; I_{ph}: photocurrent; R_{st}: row reset signal; R_d: row readout signal. The latch will be locked when clock signal (clk) is at a high level. (c) Principle of spike triggering and spike coding. The pixel intensity is encoded as 1 at the time when the flip signal (spike) is triggered and 0 otherwise. (d) Timing diagram of the spike stream. (e) Microphotograph of the VidarOne chip. Each fabricated block corresponds to the components in (a). (f) Image of the packaged VidarOne chip. (g) The spike camera system includes a visual information acquisition module composed of an industrial camera lens and a VidarOne chip, a high-speed sensing module implemented by a field-programmable gate array chip, and a real-time visual computing module implemented by a desktop workstation.

spike streams similar to the spike camera (see Section 2). In addition, this simulator provides color images by simulating the red–green–blue (RGB) channels of the pixel. Here, we construct the scenes of "PKU flying ball" and "PKU coin" with Blender and generate spike streams with Spike-Sim. The first and second rows of Fig. 4 (d) present the reference images generated by an ordinary camera (f = 30 Hz) for the two scenes and the simulated spikes generated by Spike-Sim, respectively. The TFW and TFI reconstruction results are shown in the third and fourth rows. The details in the reference images are fuzzy, while the images reconstructed based on the spike streams generated by Spike-Sim show more texture details.

3.4. Super vision system with SNNs

Here, we show that super vision can be achieved by combining the speed of the machine and the mechanism of biological vision. We propose a super vision system for high-speed moving object detection, tracking, prediction, and recognition based on SNNs that is 1000 times faster than the human vision system (Fig. 5(a)). Realizing these functions involves three major challenges: first, removing spikes generated from the background/static part of the scene for subsequent high-level visual tasks; second, detecting and smoothly tracking high-speed moving objects as well as predicting the trajectory; and third, recognizing the tracked objects. To accomplish these tasks, we propose a dynamic connection gate with STP for filtering spatiotemporal spike sequences, a locally connected SNN for object detection and tracking, a CANN for prediction, and a three-layer fully connected SNN for object recognition.

The detailed structure is shown in Figs. 5(b)–(e). As the SCHSSD camera generates spike streams with a fixed frequency for the background/static part of the scene, which will hamper the

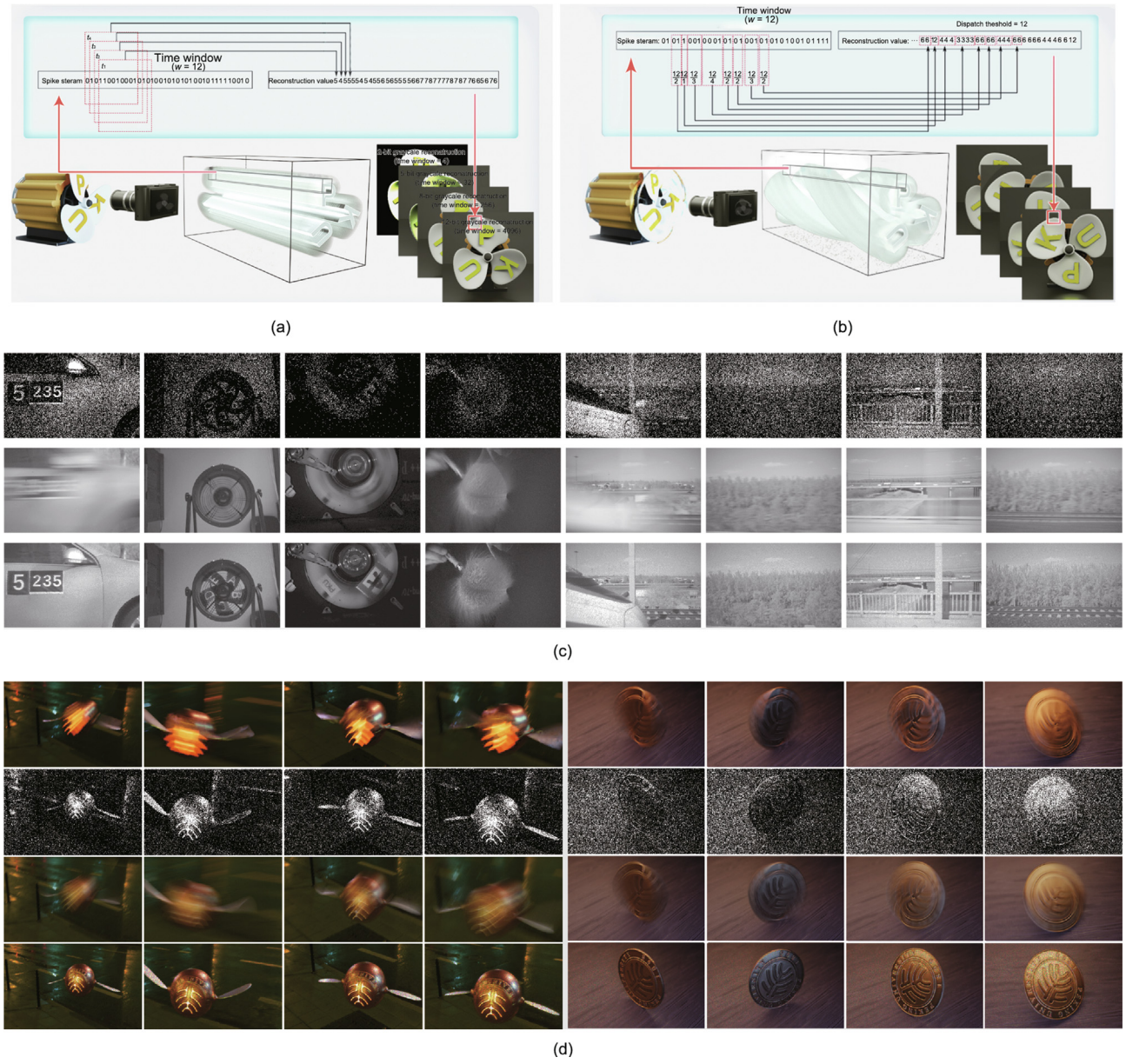

Fig. 4. Texture reconstruction with the spike camera. (a) Illustration of the TFW method. This method uses the principle that the scene radiance is directly proportional to the spike count. The light blue rectangle represents the spike streams of a pixel and the corresponding reconstructed grayscale value. The TFW method can reconstruct the texture with a free dynamic range by resizing the time window to different widths and collecting different numbers of spikes (see the four frames on the right). (b) Illustration of the TFI method. This method uses the principle that the scene radiance is inversely proportional to the interspike interval. The TFI method applies to high-speed moving scenes. (c) Reconstruction results for the SCHSSD. The three rows represent the raw spikes from the spike camera, texture reconstruction by TFW, and texture reconstruction by TFI. (d) Reconstruction results for two scenes constructed with a Blender. Scene 1: PKU flying ball (inspired by the "golden snitch" Quidditch game ball from the Harry Potter Series). The wings of the flying ball flap at night and enter the field of view of the spike camera from far to near. Scene 2; PKU coin. A gold coin with the Peking University logo rotates on a wooden desktop and eventually stops. Both scenes consider the slight movement of the camera.

subsequent high-level visual tasks, the dynamic connection gate based on STP is introduced here to filter spikes (Fig. 5(b), see Section 2). The gate closes when the input spike streams have a fixed frequency (corresponding to background or static objects) and opens when the spike frequency changes (corresponding to moving objects); thus, only the spike streams generated by the moving objects are retained. The neurons in the filter layer send excitatory PSPs (EPSPs) to spatially adjacent neurons in the detection layer, where all the neurons fire spikes according to the LIF model (Fig. 5(c)). Each moving object is found by detecting the connected area of the firing neurons, while in the tracking layer, different moving objects are associated by comparing the position or topology similarity of the moving neurons at the previous time with that at the current time. The next layer is a CANN (Fig. 5(d)), which is used to predict the trajectory by adding negative feedback to the neuronal dynamics. It can track a moving object anticipatively with an approximately constant leading time (see Section 2). The recognition network is a multilayer fully connected SNN (Fig. 5(e)). The network is trained with the BP-STDP learning rule (see Section 2), and the recognition result is determined by the firing rate of the neurons in the last layer.

3.5. Demonstration of the utility of the spike camera and the super vision system

To demonstrate the utility of the spike camera and the super vision system, we design experiments of auxiliary referee and target pointing systems. Fig. 6(a) illustrates the auxiliary referee

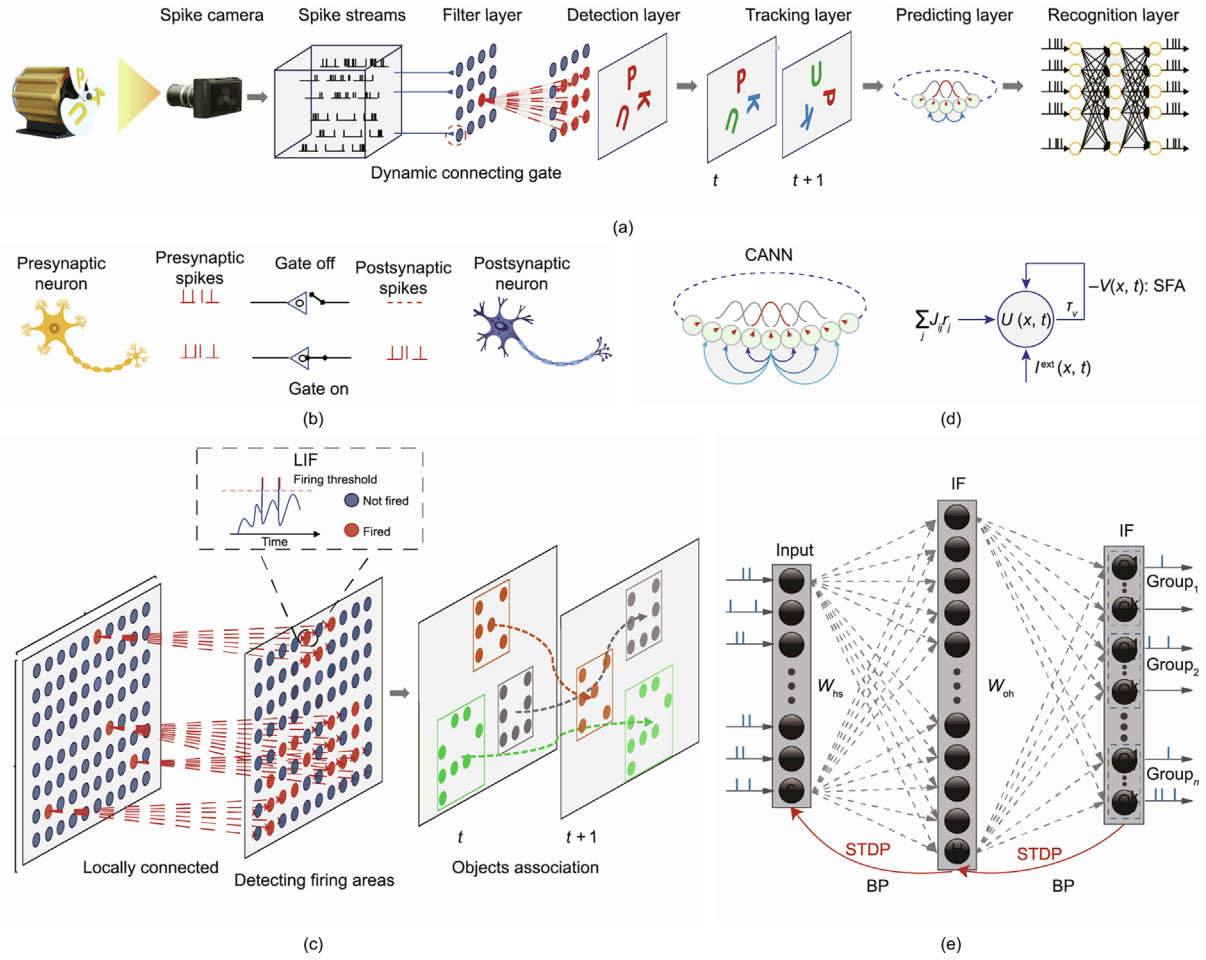

Fig. 5. Super vision system. (a) Framework of high-speed moving object detection, tracking, prediction and recognition based on SNNs. (b) Dynamic connection gate with STP for the spatiotemporal spike sequence filter, it removes the spike streams that have a fixed frequency. (c) Locally connected SNN for object detection and tracking. (d) CANN for predicting the trajectory, the right of the figure is the dynamics of a single neuron, where the neuron receives a recurrent input $\sum_j J_{ij} r_j$ from other neurons, an external input $I^{ext}(x,t)$ containing the stimulus information and a negative current $-V(x,t)$ representing the spike–frequency–adaption (SFA) effect; the feedback of SFA is effectively delayed by time τ_v; $U(x,t)$: the synaptic input at time t to neuron at x. (e) Three-layer fully connected SNN trained for object recognition. IF: integrate-and-fire neuron; W: weight.

scene, in which we make use of a table tennis ball machine to launch a ball to simulate games such as tennis and badminton. The problem is to determine whether the ball is in or out of bounds (white line). As making a judgment by the human eye when the ball drop location is near the bounds is difficult, an eagle eye system is often used in the game. In addition to being expensive, the eagle eye system cannot record the moment when the ball hits the ground. Generally, it is calculated from the movement trajectory, which may cause disputes with a referee. In contrast, the spike camera has the full-time imaging ability to record the entire process of ball landing, thus enabling the referee to determine the ball drop location (Fig. 6(b)).

Moreover, a target pointing system is presented to demonstrate that high-speed vision can be achieved by combining the spike camera and the super vision system (Fig. 6(c)). The spike camera is set in front of a high-speed rotating fan (approximately 2400 rpm) with three characters ("P," "K," and "U") pasted on its blades. One can choose one of the characters as the pointing target, and the laser needs to fire and hit the photographic paper at the top of the character. Solving this problem requires three steps: first, detecting and tracking all the moving objects in the scene; second, recognizing all the moving objects and determining the position of the character given in advance; and third, predicting the trajectory of the character and controlling the laser to hit the target. The proposed super vision system (Fig. 5) is used to accomplish this task. Fig. 6(d) illustrates a comparison of the fan before and after laser hits. This demonstration provides an appropriate method to evaluate the performance of the spike camera and the super vision system. Fig. 6(e) visualizes the output spike train in response to different characters. Fig. 6(f) illustrates the detection and tracking performance for the three characters, from which one can find that the network can detect all the moving objects and track them smoothly. The spike camera and the super vision system can detect, track, and recognize the fan moving with a linear velocity of 30 m·s^{-1} within 0.75 m in real time (see Section 2). According to the central perspective principle, this system can detect, track, and recognize an aircraft flying at the speed of sound within 10 m. Moreover, it can detect, track, and recognize a high-speed moving object at Mach 100 within 1 km (Fig. 6(g)).

3.6. Application prospects

The vidar camera is capable of capturing fast object movements. This camera has a high-speed mode with functions similar to the human eye and better performance than the human eye. This is not possible with traditional frame-based cameras due to the significant information loss between frames. By increasing the frame rate, some high-speed cameras, such as Phantom cameras, have mitigated this problem, but they require specialized sensors and shutters that are highly expensive. In contrast, the spike camera

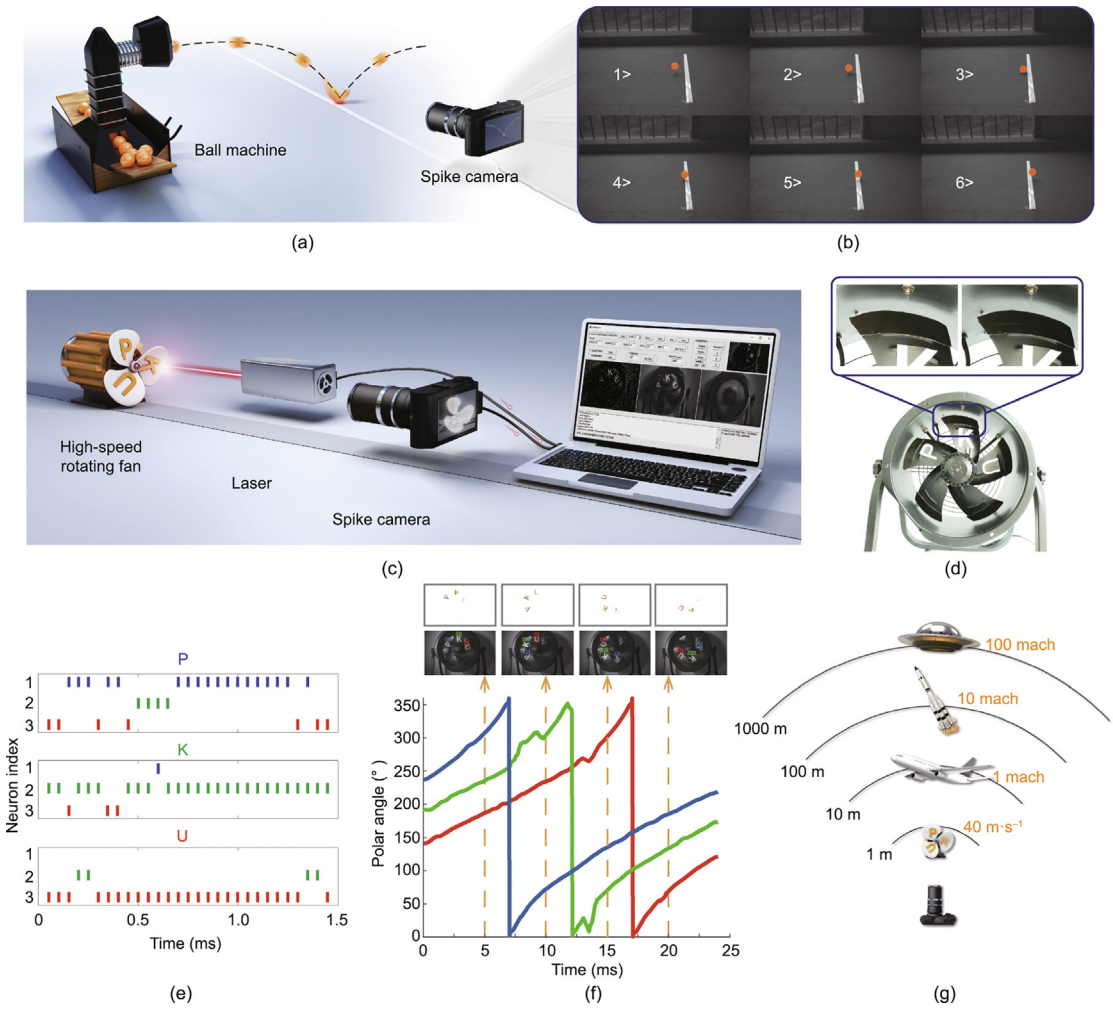

Fig. 6. Demonstration of the utility of the vidar camera and the supervision system with auxiliary referee and target pointing systems. (a) Illustration of the auxiliary referee task, where the spike camera is used to determine whether the ball is in or out of bounds. (b) The spike camera can capture the entire process of the ball landing. Here, the speed of the table tennis ball is approximately 100 $km \cdot h^{-1}$. Note that the ball and boundary are colored for emphasis, and only 6 of 170 frames are shown here. (c) Illustration of the target pointing system. The laser needs to hit the laser printing paper at the top of the known character on a high-speed rotating fan. (d) Comparison of a fan before and after laser hits. The laser sends 64 pulses and hits the predetermined character "K." (e) SNN recognition test. The neuron in the output layer of the recognition SNN corresponding to the correct category produces the most spikes. (f) Multiobject tracking by detection. The *y* axis shows the polar angle of each object's position center point relative to the center of the fan. The SNNs can obtain a mask of each character and return its bounding box in real time. The mask and bounding box are colored by object membership. (g) Evaluation of the performance of the vidar camera and the super vision system. 1 mach = 340.3 $m \cdot s^{-1}$.

is far more cost-effective because it is built from conventional CCDs via regular semiconductor manufacturing processes. Thus, it can be widely used in daily life, such as with mobile phones and cameras.

Another advantage of the spike camera over traditional cameras is that it offers a more flexible image acquisition method. The spike camera can reconstruct the image at any given moment with considerable flexibility in the dynamic range. Distinct from another retina-inspired camera, the dynamic vision sensor [20–22], whose photosensitive units only generate events when the brightness change exceeds a threshold, each photosensitive unit of the spike camera keeps capturing photons independently and generates a spike when the accumulated intensity exceeds the given threshold. Therefore, the scene radiance at each sampling position is effectively recorded by the spike camera. We believe that this system will create its own niche in surveillance systems, with applications to dynamic face recognition, fingerprint recognition, and palm print recognition.

The spike camera is inspired by the neuronal circuitry structure and information processing mechanism in the primate fovea. It converts light signals into electrical signals, yielding spike trains as output that can be naturally processed by SNNs. Given that SNNs are efficient and effective in visual perception and cognitive tasks, we expect that the combination of the spike camera and the super vision system based on SNNs will provide plentiful utilities for both fundamental research questions and practical applications, such as object detection, recognition, and tracking at electrical speeds.

4. Conclusions

By replacing video with vform, spike cameras will bring camera development back on the right track, realizing the technological potential of optoelectronic technology that has been suppressed for decades and replacing traditional video cameras in almost all fields, especially for high-speed scenes, thus triggering a revolution in the camera field.

The essence of vform is a spike stream that characterizes the process of optical temporal and spatial changes, which is a natural input of SNNs. Vform, as a new generation of the eye of machine vision, will play an important role in the era of artificial intelligence.

Acknowledgments

This work was supported by projects of the National Natural Science Foundation of China (61425025) and the Beijing Municipal Science and Technology Project (Z151100000915070 and Z171100000117008).

Compliance with ethics guidelines

Tiejun Huang, Yajing Zheng, Zhaofei Yu, Rui Chen, Yuan Li, Ruiqin Xiong, Lei Ma, Junwei Zhao, Siwei Dong, Lin Zhu, Jianing Li, Shanshan Jia, Yihua Fu, Boxin Shi, Si Wu, and Yonghong Tian declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Haykin S, Van B. Signals and systems. New Jersey: John Wiley & Sons; 2007.
[2] Chakravorty P. What is a signal? IEEE Signal Process Mag 2018;35(5):175–7.
[3] Stump D. Digital cinematography: fundamentals, tools, techniques, and workflows. Boca Raton: CRC Press; 2014.
[4] Itatani J, Quéré F, Yudin GL, Ivanov MY, Krausz F, Corkum PB. Attosecond streak camera. Phys Rev Lett 2002;88(17):173903.
[5] Bradley DK, Bell PM, Landen OL, Kilkenny JD, Oertel J. Development and characterization of a pair of 30–40 ps x-ray framing cameras. Rev Sci Instrum 1995;66(1):716–8.
[6] Wässle H. Parallel processing in the mammalian retina. Nat Rev Neurosci 2004;5(10):747–57.
[7] Masland RH. The neuronal organization of the retina. Neuron 2012;76 (2):266–80.
[8] Litwiller D. CCD vs CMOS. Photon Spectra 2001;35:154–8.
[9] Lamb TD, Pugh EN. Phototransduction, dark adaptation, and rhodopsin regeneration the proctor lecture. Invest Ophthalmol Vis Sci 2006;47(12):5137–52.
[10] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.
[11] Maass W. Networks of spiking neurons: the third generation of neural network models. Neural Netw 1997;10(9):1659–71.
[12] Roy K, Jaiswal A, Panda P. Towards spike-based machine intelligence with neuromorphic computing. Nature 2019;575(7784):607–17.
[13] Marr D, Poggio T, Ullman S. Vision: a computational investigation into the human representation and processing of visual information. Cambridge: MIT Press; 2010.
[14] Palmer SE. Vision science: photons to phenomenology. Cambridge: MIT Press; 1999.
[15] Li Z. Understanding vision: theory, model, and data. New York City: Oxford University Press; 2014.
[16] Davies ER. Computer and machine vision: theory, algorithm, practicalities. 4th ed. London: Academic Press; 2012.
[17] Sonka M, Hlavac V, Boyle R. Image processing, analysis, and machine vision. 4th ed. Stamford: Cengage Learning; 2015.
[18] Medathati NVK, Neumann H, Masson GS, Kornprobst P. Bio-inspired computer vision: towards a synergistic approach of artificial and biological vision. Comput Vis Image Underst 2016;150:1–30.
[19] Tsodyks MV, Markram H. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. Proc Natl Acad Sci USA 1997;94(2):719–23.
[20] Tsodyks M, Pawelzik K, Markram H. Neural networks with dynamic synapses. Neural Comput 1998;10(4):821–35.
[21] Costa RP, Sjöström PJ, van Rossum MCW. Probabilistic inference of short-term synaptic plasticity in neocortical microcircuits. Front Comput Neurosci 2013;7:75.
[22] Mi Y, Fung CA, Wong KM, Wu S. Spike frequency adaptation implements anticipative tracking in continuous attractor neural networks. Adv Neural Inf Process Syst 2014;27:505–13.
[23] Tavanaei A, Maida A. BP-STDP: approximating backpropagation using spike timing dependent plasticity. Neurocomputing 2019;330:39–47.
[24] Song S, Miller KD, Abbott LF. Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. Nat Neurosci 2000;3(9):919–26.
[25] Rumsey CC, Abbott LF. Synaptic equalization by anti-STDP. Neurocomputing 2004;58:359–61.
[26] Adelson EH, Bergen JR. The plenoptic function and the elements of early vision. In: Landy MS, Movshon JA, editors. Computational models of visual processing. Cambridge: MIT Press; 1991.