Retina-like Visual Image Reconstruction via Spiking Neural Model

Lin Zhu[1,2] Siwei Dong[1] JianingLi[1,2] Tiejun Huang[1] Yonghong Tian[1,2,*]
Peking University[1] Pengcheng Laboratory[2]
{linzhu,swdong,lijianing,tjhuang,yhtian}@pku.edu.cn

Abstract

The high-sensitivity vision of primates, including humans, is mediated by a small retinal region called the fovea. As a novel bio-inspired vision sensor, spike camera mimics the fovea to record the nature scenes by continuous-time spikes instead of frame-based manner. However, reconstructing visual images from the spikes remains to be a challenge. In this paper, we design a retina-like visual image reconstruction framework, which is flexible in reconstructing full texture of natural scenes from the totally new spike data. Specifically, the proposed architecture consists of motion local excitation layer, spike refining layer and visual reconstruction layer motivated by bio-realistic leaky integrate and fire (LIF) neurons and synapse connection with spike-timing-dependent plasticity (STDP) rules. This approach may represent a major shift from conventional frame-based vision to the continuous-time retina-like vision, owning to the advantages of high temporal resolution and low power consumption. To test the performance, a spike dataset is constructed which is recorded by the spike camera. The experimental results show that the proposed approach is extremely effective in reconstructing the visual image in both normal and high speed scenes, while achieving high dynamic range and high image quality.

1. Introduction

Autonomous driving, wearable computing, unmanned aerial vehicles, are typical emerging real-time applications which require rapid reaction in vision processing [19]. Conventional cameras compress the video data in the exposure time into one frame, and the temporal changes in that time will be lost [11]. When performing image analysis tasks such as detecting or tracking an object, these consecutive frames have to be compared to recover temporal changes, which is computationally expensive and is difficult to achieve satisfactory results [21].

If we turn attention to the human vision, the visual sampling is quite different from that of a digital camera. There is no concept of frames or pictures in human vision. Although the mechanism of human vision is too complicated to be fully understood, the physical structures and the signal processing in human retina give us some hints and inspirations. Among them, the dynamic vision sensor (DVS) is the most well-recognized [8, 1]. In DVS, each pixel responds independently to the changes of luminance intensity by generating asynchronous spikes. This mechanism is similar to the periphery of the retina, which is sensitive only to moving objects. The temporal redundancy of the output spikes is natively reduced, however, it is not able to reconstruct the visual images as the conventional camera does. Although there are some hybrid sensors combing DVS and conventional image sensor (DAVIS) [5], or adding an extra photo-measurement circuit (ATIS [23], CeleX [13]), there exists motion mismatch since the difference of the sampling time resolution.

To solve the problem of capturing visual texture while maintaining the continuous-time signal, researchers designed a class of time-based sensors to make each pixel mimics the behaviour of an integrate-and-fire neuron and works asynchronously [2, 6, 16]. Instead of choosing a fixed integration time for all pixels like a conventional camera, the time-based sensor ensures that each pixel selects its own optimal integration time to achieve a high dynamic range and an improved signal-to-noise ratio. This kind of sensor enables the reconstruction of visual textures in a frame-free manner. Using a time window or the inter-spike interval, the image texture can be reconstructed [6].

Recently, [28] proposed a fovea-like sampling method (FSM) which falls into the category of time-based sensors. Compared to previous time-based sensors, this sensor namely spike camera is with high spatial (250×400) and temporal resolutions (40000 Hz), which is suitable to deal with high-speed vision tasks [28]. However, the previous reconstruction algorithms [6, 28] will suffer the problem of low contrast or blur in complex environments. Therefore, how to flexibly use the time-continuous spike information is a key problem of high-quality image reconstruction.

In this paper, we propose a new retina-like visual im-

*Corresponding author.

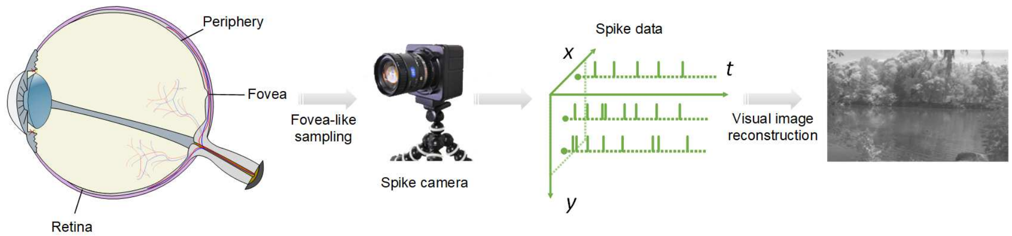

Figure 1. **The spike camera based on fovea-like sampling and visual image reconstruction.**

age reconstruction framework, as shown in Fig 1. The main contributions of this paper includes: 1) We propose a three-layer spiking neural model which relies on a combination of biologically plausible mechanisms. Using time-continuous spike data, our method can reconstruct images at any sampling moment, and retain the details of high-speed motion and static background simultaneously. 2) We propose a dynamic neuron extraction model to distinguish the neuron states (dynamic or static) in an incremental way, which is useful for reconstructing high quality high-speed motion scenes. 3) We construct a new spike dataset for evaluating the reconstruction method, and make these available to the research community.

2. Spike Data Analysis

2.1. Spike Data Representation

In FSM, the intensity of light is converted into voltage by the photoreceptor [28]. Once the voltage reaches a predefined threshold, a one-bit spike is outputted and a signal to reset the integrator is dispatched at the same time. This process is quite similar to the integrate-and-fire neuron. Different luminance stimuli I leads to a different spike firing rate, the output and the reset are triggered asynchronously among various pixels. Typically, the brighter the light, the faster the firing speed:

$$\int I dt \geq \phi \tag{1}$$

The raw data from the spike camera is a three-dimensional spike array D. The spike camera only cares about the integration of the luminance intensity and fires spikes in an ultra high frequency. At each sampling moment, if a spike is just fired, a digital signal "1" (i.e. a spike) is outputted, otherwise "0" is generated. We define $S_{i,j}(t) \in \{0, 1\}$ to represent the spike firing status of pixel (i, j) at the moment t. For simplicity, we use **spike plane** to represent the spike signal outputted by all pixel at a certain moment, while the time-continuous spike signals generated by a certain pixel is called a **spike train** (see Fig 5 (a)).

2.2. Spike Data Distribution

The integrator has a predefined capacity which is also known as the spike firing threshold ϕ. If the integrator is filled, it will be reset and fires a spike. Due to the variation of the light, the duration of filling the integrator from empty to fulfilled is not constant. Microscopically, a spike is generated means a fixed number of photons have been recorded. We define $N(t, \delta)$ as the number of photons arrived at the photoreceptor within the time interval $[t, t + \delta)$, and $R(t, \delta)$ as the number of photons recorded actually in the same period. However, the dead time τ between two consecutive photon arrivals makes $N(t, \delta)$ and $R(t, \delta)$ are not equal. If the former arrival is recorded at time t, any latter photon arrivals during $(t, t + \tau]$ will not be recorded.

In fact, the photon record process $R(t, \delta)$ can be seemed as a renewal process, which involves recurrent patterns after each of which the process starts from scratch. The photon arrival process is usually assumed to be a homogeneous Poisson process. It is parameterized by a single scalar λ which gives the mean rate of the photon arrivals. If the waiting time between one renewal and the next has ensemble mean and variance, the photo recording process with dead time τ is asymptotically Gaussian distributed [7]:

$$E \sim \frac{\lambda\delta}{1 + \lambda\tau}, Var \sim \frac{\lambda\delta}{(1 + \lambda\tau)^3} \tag{2}$$

To validate the model, we record several spike sequences using the spike camera under various light conditions. We assume that the record of n photons will reach the dispatch threshold ϕ and generate a spike. If the spike firing time is denoted as t_i, the inter-spike interval is $t_{isi} = t_i - t_{i-1}$. As shown in Fig 2, the blocks with different grayscale values represent the luminance intensities, which indicates that larger intensities lead to higher spike firing rates and shorter inter-spike intervals (ISIs). The RMSE shows that the inter-spike interval distribution histogram can be well fitted by the approximate Gaussian distribution.

Based on the above, we are able to model the ISI distribution of a certain intensity by a Gaussian distribution. In Sec 3.2, a dynamic neuron extraction model is proposed to extract the spike signal representing moving object according to the ISI distribution.

3. Spiking Neural Model

To address the challenge of visual image reconstruction from the spike data, we propose a novel spike neural mod-

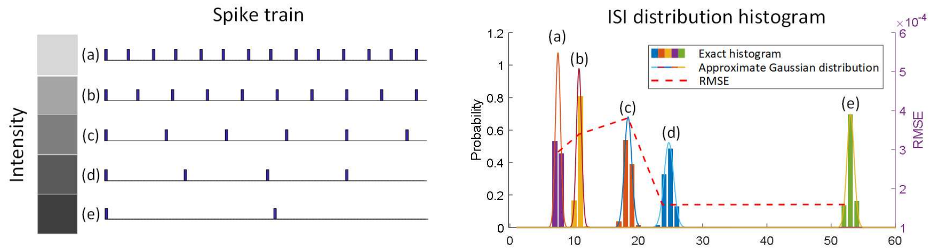

Figure 2. **The spike data distribution under different intensities.** Left: The spike train generated by different light intensity. Larger intensities lead to higher spike firing rates and shorter inter-spike intervals. Right: the exact ISI distribution histogram, approximate Gaussian distribution and their Root Mean Squard Error (RMSE). The RMSE shows that the approximate Gaussian distribution can well fit the real ISI distribution.

el based on the inspiration from the biological neural dynamics and the adaptation. In biological neural systems, a neuron receiving the stimuli and firing a spike can be abstracted as a leaky integrate-and-fire (LIF) model [14]. In LIF model, the membrane potential $V(t)$ is governed by the following differential equation:

$$\tau_m \frac{dV}{dt} = -(V(t) - V_{rest}) + RI(t) \qquad (3)$$

where V_{rest} is the rest potential, $I(t)$ is the total synaptic current, R is the membrane resistance and τ_m is the membrane time constant. In the absence of input, the membrane potential decays exponentially to its resting. With the input spikes, each input yields the input potential onto the membrane potential $V(t)$. Each time the membrane potential hits the threshold, the membrane potential $V(t)$ is reset to V_{rest} and a spike is fired [14]. The refractory period in a neuron occurs after one output spike, which is quite different in various neurons.

In addition, there exists multiple adaptation mechanisms in biological neuron. The neuron is adaptively adjusted according to the input spike characteristics. For instance, the synapse plasticity [3, 4] modulates the efficiency of neural connections by its weight, while the membrane potential thresholds (spike firing thresholds) of various neurons are different to adapt to different stimuli. The dynamic threshold [9] allows that more frequent stimuli may lead to greater thresholds, and vice versa.

Based on the inspiration from biological neurons, we propose a spiking neural model to solve the problem of image reconstruction. The model shown in Fig 3 includes motion local excitation layer, spike refining layer and visual reconstruction layer. The motion local excitation layer receives spikes input and marks the motion state of neurons (static or dynamic). Then the spike refining layer adjusts the refractory period of each neuron according to its motion state, which acts as a temporal filter. The last visual reconstruction layer adopts the adaptation mechanism widely existed in biological neurons, and maps the dynamic threshold of each neuron into a grayscale image as the output.

The overall architecture of the spiking neural model is illustrated in Fig 3 and in more detail in Fig 5. The architecture of the spiking neural model is described in Sec 3.1. The dynamic neuron extraction based on graph cuts is shown in Sec 3.2. Sec 3.3 presents the synapse connection, while the visual image reconstruction is introduced in Sec 3.4.

3.1. Model Architecture

Motion Local Excitation Layer The motion local excitation layer operates on the input spike data and outputs spike train with binary marks (dynamic or static). The purpose is to distinguish the neuron state according to the input spike data. Each neuron in this layer is connected to the input spike data in one-to-one connections. In this layer, the input spike data is modelled as a motion confidence matrix according to the historical firing distribution. Then, for current moment, the neuron states can be abstracted into the first-order Markov Random Field with binary labels [12], and a motion local extractor based on the graph cuts is performed. In this way, each output spike has a dynamic or static mark to distinguish the state of the neuron. The detail will be described in Sec 3.2.

Spike Refining Layer To model neuronal dynamics, the LIF model is introduced in spike refine layer. In this layer, the input spikes are filtered to keep the fast response to the motion while removing the noise. The size of this layer is the same as the motion local excitation layer, each input is fed to one neuron in this layer. In order to rapidly respond to the motion, we set the threshold voltage to a very small value. To eliminate the noise, according to the mark given by the motion local excitation layer, a relatively long refractory period should be set if the current input spike is marked static; otherwise, if the spike comes from a dynamic neuron, a relatively short refractory period should be set. By the above mechanism, the spike refining layer can significantly eliminate the noise, and mitigate the over-exposure by reducing the firing rate. Meanwhile, this layer retains as much detail as possible to maintain the fidelity of dynamic spike while preserved high dynamic range.

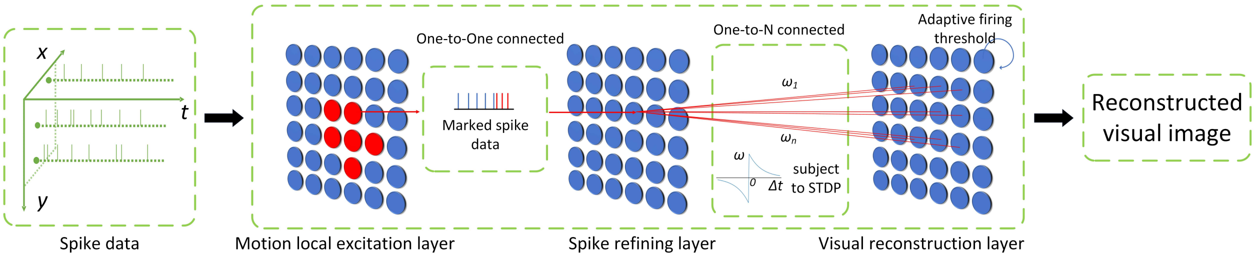

Figure 3. **The overall architecture of spiking neural model.**

Visual Reconstruction Layer As in the previous layer, the neurons in this layer are LIF neurons. Each neuron receives the spike train from the spike refining layer, and the neuron and the synaptic connection make various adaptive adjustments. The visual image is reconstructed according to the state of the neuron. The neurons of this layer are connected in a one-to-many fashion to that of spike refining layer. These neurons have no refractory period. STDP learning rule [4] is performed between these two layers to learn to adjust the spike firing rate (see Sec 3.3). In addition, to achieve the homeostasis of the system, threshold adaptation [9] is introduced in this layer. According to the statistics of the neuron state and dynamic threshold, high quality visual images including dynamic and static scenes can be reconstructed simultaneously (see Sec 3.4).

3.2. Dynamic Neuron Extraction

In this section, we propose a dynamic neuron extraction model to mark input spikes as dynamic or static. As analyzed in Sec 2.2, the ISI is equal to the time of a fixed number of photons have been recorded, which is proportional to the firing threshold ϕ of the integrator. For a constant photon arrival rate λ, the ISI distribution has an unimodal and symmetric distribution which approaches a Gaussian distribution. The region visited by a moving object has a dissimilar ISI distribution from that driven by static region. Therefore, we associate the ISI of each neuron with a Gaussian probability model with mean μ and covariance σ:

$$\text{ISI} \sim \mathcal{N}(\mu, \sigma) \tag{4}$$

Assuming that each neuron corresponds to a coordinate $(i, j), i \in [1, m], j \in [1, n]$, where m and n are the resolution of spike camera. The ij-th neuron is denoted as ij, we evaluate all the moments against their corresponding ISI distribution models and obtain the confidence map $C_{ij}(t) \in \mathbb{R}^{T \times 1}$, which corresponds to the confidence of the location belong to static region. Therefore, the motion confidence of each neuron at different moment is:

$$O = \begin{bmatrix} C_{11} & \cdots & C_{1n} \\ \vdots & \ddots & \vdots \\ C_{m1} & \cdots & C_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n \times T} \tag{5}$$

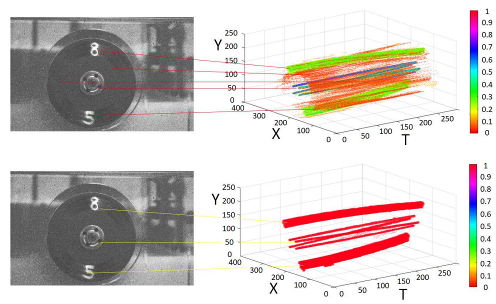

Figure 4. **The illustration of dynamic neuron extraction.** This operation is performed in motion local excitation layer. The dynamic neurons corresponding to digital and turntable center are extracted. Top: the visualized result of the motion confidence matrix O. Bottom: the dynamic neuron extraction results.

At the moment t, the mark matrix $M^t \in \{0, 1\}^{m \times n}$ is a binary matrix denoting the states of the neurons:

$$M_{ij}^t = \begin{cases} 1 & \text{if } ij \text{ belongs to motion region at moment } t \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

We use $\mathcal{P}_{M^t}(X)$ to represent the orthogonal projection of a matrix X onto the linear space of matrices supported by M^t,

$$\mathcal{P}_{M^t}(X)(i,j) = \begin{cases} 0 & \text{if } M_{ij}^t = 0 \\ X_{ij} & \text{if } M_{ij}^t = 1 \end{cases} \tag{7}$$

and $\mathcal{P}_{M^{t\perp}}(X)$ to be its complementary projection, i.e., $\mathcal{P}_{M^t}(X) + \mathcal{P}_{M^{t\perp}}(X) = X$.

The binary matrix M^t can be naturally modeled by a Markov Random Field (MRF) [12]. Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices denoting all $m \times n$ neurons and $\mathcal{E}$ is the set of edges connecting spatially neighboring neurons. According to the Ising model [17], the energy of M can be represented as:

$$\sum_{ij \in \mathcal{V}} u_{ij}(M_{ij}) + \sum_{(ij,kl) \in \mathcal{E}} \lambda_{ij,kl} \left| M_{ij} - M_{kl} \right| \tag{8}$$

where

$$u_{ij}(M_{ij}) = \begin{cases} \lambda_{ij} & \text{if } M_{ij} = 1 \\ 0 & \text{if } M_{ij} = 0 \end{cases} \tag{9}$$

where $\lambda_{ij,kl}$ controls the strength of dependency between M_{ij} and M_{kl}, and λ_{ij} controls the sparsity of $M_{ij} = 1$.

Since O^t denotes the motion confidence, we assume that $O^t = M^t + N^t$, where N^t denotes the noise. For the static neurons of $M_{ij}^t = 0$, the noise $N^t = \mathcal{P}_{M^{t\perp}}(O^t)$ should be minimized. Combining the above motion model and noise model, we propose the following optimization:

$$\min_{M_{ij}^t \in 0,1} \frac{1}{2} \sum_{ij} \mathcal{P}_{M^{t\perp}}(O_{ij}^t)^2 + \sum_{ij \in \mathcal{V}} u_{ij}(M_{ij}^t) + \sum_{(ij,kl) \in \mathcal{E}} \lambda_{ij,kl} \left| M_{ij}^t - M_{kl}^t \right| \tag{10}$$

For simplicity, we set $\lambda_{ij} = \alpha$ and $\lambda_{ij,kl} = \beta$. Eq. 10 can be rewritten as follows:

$$\min_{M_{ij}^t \in 0,1} \sum_{ij} (\alpha - \frac{1}{2}(O_{ij}^t)^2) M_{ij}^t + \beta \left\| E\text{vec}(M^t) \right\|_1 + c \tag{11}$$

where $c = \frac{1}{2}\sum_{ij}(O_{ij}^t)^2$, E is the node-edge incidence matrix of $\mathcal{G}$. Eq. 11 is the standard form of the first-order MRFs with binary labels, which can be solved exactly by graph cuts [15].

The illustration of dynamic neuron extraction is shown in Fig 4. The proposed model is sufficient to distinguish the neuron state. When a new set of spikes arrives, the confidence matrix is updated incrementally, and the graph cuts can be operated for each moment separately.

3.3. Synapse Connection

Synapse is a structure that permits a neuron to pass a voltage signal to another neuron according to their weights. In applications of spiking neural network (SNN), the weights can be trained following a learning rule or set to constant values. Synaptic plasticity is the basic mechanism of underlying learning in biological networks [3]. It is defined as the ability to modulate the efficiency (also known as weight) of neural connections. In the biological vision system, information coding is established in an unsupervised way. Among the SNN learning rules, spike timing dependent plasticity (STDP) [3] is the most popular one. With STDP, if the presynaptic spike to a neuron tends to shortly before it fires, then the synaptic weight is made stronger (Long-Term Potentiation, LTP); whereas if an input spike tends to occur immediately after an output spike, then that particular input is made somewhat weaker (Long-Term Depression, LTD). In this work, biological STDP [4] is used to learn the rule of spike firing of static neurons, which is defined as:

$$\Delta\omega = \sum_{t_{pre}} \sum_{t_{post}} W(t_{post} - t_{pre}) \tag{12}$$

where $\Delta\omega$ is the weight variation, t_{pre} and t_{post} denote the firing time of the input neuron (presynaptic spike time) and output neuron (postsynaptic spike time), respectively. The function W is defined as:

$$W(\Delta t) = \begin{cases} A_{pre} \exp\left(-\Delta t / \tau_{pre}\right) & \text{if} \quad \Delta t > 0 \\ A_{post} \exp\left(\Delta t / \tau_{post}\right) & \text{if} \quad \Delta t < 0 \end{cases} \tag{13}$$

where the parameters A_{pre} and A_{post} depend on the current value of the synaptic weight ω.

The neurons between two adjacent layers can be connected in different ways. As mentioned above, between the second and third layers, we use a one-to-many manner to connect their neurons. We assume that neurons are distributed in a regular grid and that the distance between adjacent neurons is a constant. A presynaptic neuron will connect to multiple postsynaptic neurons if the following conditions are met:

$$\sqrt{(x_{pre} - x_{post})^2 + (y_{pre} - y_{post})^2} < R \tag{14}$$

where x and y are the coordinates of neurons, R is the connection range. And the initial weights for each synapse are obtained as following:

$$\omega = \exp(-k\sqrt{(x_{pre} - x_{post})^2 + (y_{pre} - y_{post})^2}/R) \tag{15}$$

where k controls the weight distribution. The initial weights are adaptively updated at the arrival of each spike according to STDP learning rules.

3.4. Visual Image Reconstruction

In this section, we reconstruct visual information suitable for human viewing according to the state of neuron. The inhomogeneity of the input leads to different firing rates of the excitatory neuron. In order to ensure that the neurons adapt to input spike train, we hope that all neurons will have approximately equal firing rates. To this end, a common method to adapt thresholds is to use leaky adaptive thresholds [9]: when a neuron fires a spike, a dynamic adjusting is performed for threshold to adapt the firing rate to prevent it from firing too often. The more frequently a neuron fires spikes, the higher will be its threshold. In turn, the neuron needs more inputs to fire a spike in the near future. Similar to [10], we define the model of the dynamic threshold as:

$$\vartheta_{ij}(t) = \vartheta_0 + \int_0^\infty \theta_{ij}(s) S_{ij}^{'}(t-s) ds \tag{16}$$

where ϑ_0 is the initial threshold of neuron in the absence of spiking, and $S_{ij}^{'}$ denotes the fired spikes of this layer. The firing threshold of the neuron is increased by an amount θ_{ij} and is exponentially decaying after firing a spike. The increase amount θ_{ij} is defined as:

$$\theta_{ij}(t) = \begin{cases} \eta_0 \exp\left(\frac{-(t-t_f)}{\tau}\right) & \text{if} \quad M_{ij}^t = 0 \\ \eta_0 \exp\left(\frac{-(t-t_w)}{\tau \int_{t_w}^t S_{ij}(x) dx}\right) & \text{if} \quad M_{ij}^t = 1 \end{cases} \tag{17}$$

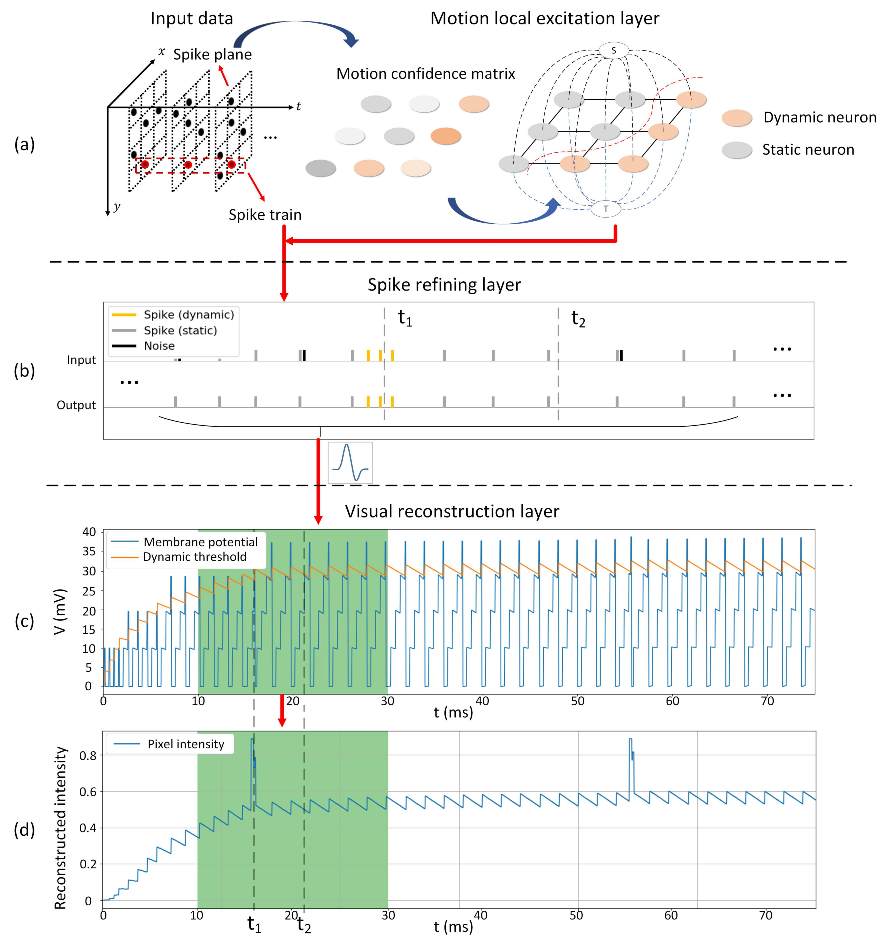

Figure 5. **The microscopic analysis of spiking neural model.** (a) The input spike data is converted to spike plane (black dashed box) and spike train (red dashed box). The spike plane connects to the motion local excitation layer, and the dynamic neurons at this moment are marked, while the spike train with the mark information input to the next layer. (b) The noise spikes are eliminated by the mechanism of the refractory period while the static and dynamic spikes are preserved. (c) Each input spikes yield a potential according to STDP, and if the accumulated membrane potential reaches the threshold, the model is adaptively adjusted to fit the input spikes. According to Eq. 18 and 19, the pixel intensity at each moment can be reconstructed (i.e. t_1: dynamic spikes, t_2: static spikes), as shown in (d).

where $S_{i,j}(t)$ is the spike train input to this layer, t_f denotes the most recent firing time , t_w is a moment before t that can be set as a constant, and τ is the time constant. In practice, we can set t_w to a small value so as to avoid threshold instability due to dynamic spike.

Finally, the grayscale value of the visual image can be estimated from the neuron state and firing threshold: if the neuron (i,j) belongs to a static neuron at moment t, in other words, $M_{i,j}^t = 0$, then the grayscale value is

$$G_{i,j,t} = \vartheta_{ij}(t) \tag{18}$$

Otherwise

$$G_{i,j,t} = \vartheta_{ij}(t^-) * \kappa(t)^\gamma \tag{19}$$

where $\vartheta_{ij}(t^-)$ denotes the convergence value of ϑ_{ij} before time t, $\kappa(t) = t^-/(t_{isi}\int_0^{t^-} S_{ij}(x)dx)$ is an adjustment parameter to reconstruct the accurate gray value of dynamic region, t_{isi} denotes the inter-spike interval corresponding to time t, and γ controls the contrast of reconstructed dynamic region.

4. Experiment

4.1. Spike Dataset

To test the proposed spiking neural model, we build a dataset including spike sequences captured by the spike camera [1]. This dataset contains eight sequences including two categories of normal speed (Class A) and high speed (Class B) scenarios. Each sequence is captured by the spike camera with 40,000 Hz sampling rate. Class A contains four sequences, of which “Office” is an indoor scene, and “Gallery”, “Lake” and “Flower” are outdoor scenes. Class

[1]https://www.pkuml.org/resources/pku-spike-recon-dataset.html.

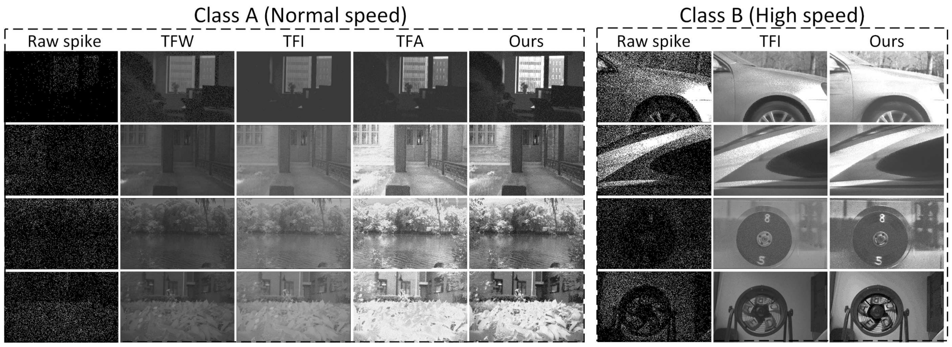

Figure 6. **The reconstruction results on Class A and B.** We compared our method with TFW, TFI and TFA [28] on Class A. Since TFP and TFA have no ability to reconstruct dynamic scenes, we only compare our methods with TFI on Class B.

Table 1. **The quantitative metrics on Class A and B.** (a higher value means better image quality)

Metric	Method	Class A (Normal speed)				Class B (High speed)				
		Office	**Gallery**	**Lake**	**Flower**	**Car**	**Train**	**Ro1**	**Ro2**	**Mean**
2-D entropy	TFW	9.12	8.68	9.45	9.22	-	-	-	-	9.12
	TFA	7.38	**12.41**	12.69	11.88	-	-	-	-	11.09
	TFI	10.01	9.85	10.51	10.01	10.71	**11.01**	10.23	9.81	10.27
	Ours	**10.38**	12.38	**12.83**	**12.63**	**10.76**	10.85	**11.72**	**11.69**	**11.66**
OG-IQA [20]	TFW	0.5729	0.4445	0.7575	**0.8959**	-	-	-	-	0.6677
	TFA	0.2737	0.6707	0.8369	0.8660	-	-	-	-	0.6618
	TFI	0.5738	0.4134	0.5110	0.8523	0.8637	0.7829	0.5602	0.5305	0.6359
	Ours	**0.5921**	**0.6879**	**0.8379**	0.8727	**0.8720**	**0.7897**	**0.6009**	**0.8105**	**0.7579**

B includes “Car”, “Train”, “Rotation1 (Ro1)” and “Rotation2 (Ro2)”. Among them, “Car” describes a car traveling at a speed of 100 km/h (kilometers per hour), while “Train” records a train with 350 km/h speed. The sequence “Rotation1” describes a disk with 2000 rpm (revolutions per minute), and the sequence “Rotation2” depicts an electric fan with 2600 rpm.

4.2. Visual Texture Reconstruction

4.2.1 Qualitative Analysis

The spiking neural model is implemented using the Brian2 neural simulator [27]. All neurons are modeled as LIF neurons. The dimension of the network is designed to fit the camera resolution of 250×400, in other words, each layer using 100,000 neurons.

To evaluate the performance of our method, we compared three methods proposed in [28], namely TFW (texture from window), TFI (texture from inter-spike interval) and TFA (adaptive texture reconstruction). The parameters of the three methods are set according to the default parameters given in original paper. Fig 6 shows the experimental results. The results of TFW and TFI have low contrasts, which makes it difficult to distinguish the details of the image. TFA improves contrast, but some regions in the image are too bright, which affects the overall visual effect of the image. Our method solves the above problems. Subjectively, the reconstructed image quality is better than the other three methods.

Class B contains four high-speed motion scenes. Since TFP and TFA have no ability to reconstruct dynamic scenes, we only compare our method to TFI on Class B. TFI is a method of image reconstruction based on instantaneous intensity, it has the ability to reconstruct high-speed motion because it conforms to the sampling principle of the camera. However, TFI only uses the information of the current moment, the historical information in temporal domain is not used. Benefit from the three-layer spiking neural model, our approach makes full use of both historical and current information. The results show that the contrast of static region is improved, while the saliency and clarity of dynamic region are enhanced.

4.2.2 Quantitative Analysis

In Table 1, we give a quantitative evaluation on the proposed dataset. Two no-reference image quality assessment metrics, two-dimensional (2-D) entropy [18] and OG-IQA [20], are employed into our experiment. 2-D entropy uses both the gray value of a pixel and its local average gray value, it measures the amount of information in the image. OG-IQA uses perceptual image features for image quality assessment and gives a score (the range is 0-1, 1 is the best). As shown in Table 1, our method achieves better than other methods in both 2-D entropy and OG-IQA metrics, this is consistent with the results of subjective observation in Fig 6.

Quantitatively, for further evaluate the high-speed scenes in Class B, we employ standard deviation (STD) and a no-reference image blur metric called CPBD [22]. The result

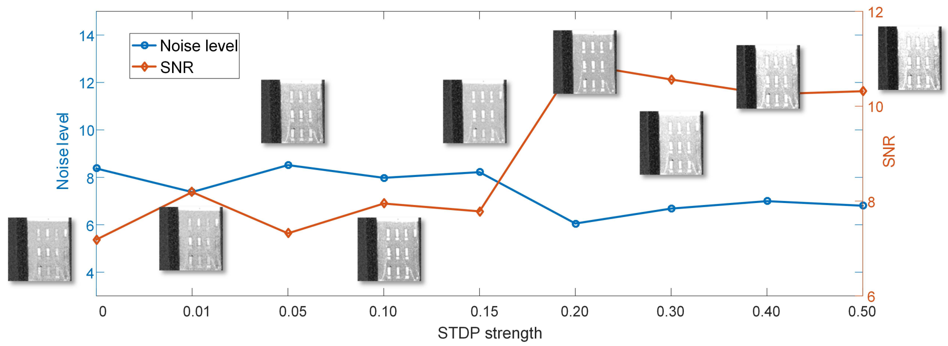

Figure 7. **The effect of STDP.** The blue line denotes the noise level estimated by [24], while the red line represents the signal-to-noise ratio. The experiment was performed on Class A. Intuitively, when the STDP strength is 0.2, the result is the best.

Table 2. **The STD and CPBD metrics on Class B.**

Metric	Method	Car	Train	Ro1	Ro2	Mean
STD	TFI	53.82	60.57	22.79	31.55	42.18
	Ours	**58.43**	**66.91**	**37.57**	**47.11**	**52.51**
CPBD	TFI	0.7797	0.9009	0.8461	0.6516	0.7946
[22]	Ours	**0.7960**	**0.9072**	**0.8482**	**0.7958**	**0.8368**

is shown in Table 2. The STD is related to the contrast of the image. Generally speaking, larger standard deviation (STD) means higher contrast. CPBD is used to measure the motion blur. Lower CPBD values mean more blur and vice versa. In summary, our method maintains sharpness while maintaining a higher contrast.

To better understand the effect of STDP, we performed an experiment on our dataset. STDP learning rule can adjust synaptic weight adaptively according to the input spike. If an irregular spike is input, which is usually caused by noise, the STDP mechanism can make its influence smaller and achieve the purpose of denoising. In the experiment, an image noise estimation method [24] is used, and signal-to-noise ratio (SNR) is obtained by standard deviation and noise. We adjust the STDP strength from 0 to 0.50, and compare the noise and SNR of the results. As shown in Fig 7, with appropriate STDP strength (0.20 gets the highest S-NR), the windows and buildings are clearer, and most of the noise is eliminated.

4.3. Comparisons with other vision sensors

Fig 8 shows the comparison of spike camera, CeleX, DAVIS240B and Huawei P30. Spike camera with the proposed reconstruction method can clearly show the detail of movement process. DVS only records the change of luminance intensity, it is very difficult to reconstruct the texture. We use two recently published DVS reconstruction methods [26] and [25] to generate visual image, but the results are unsatisfactory; while CeleX can roughly see the process of falling pages, but the shadow is serious, the edges of front and back pages can not be clearly distinguished; Huawei P30 rear camera can record videos with 60 FPS in default mode, the pages and electric fan are blurred. The video results can be found in our supplementary material.

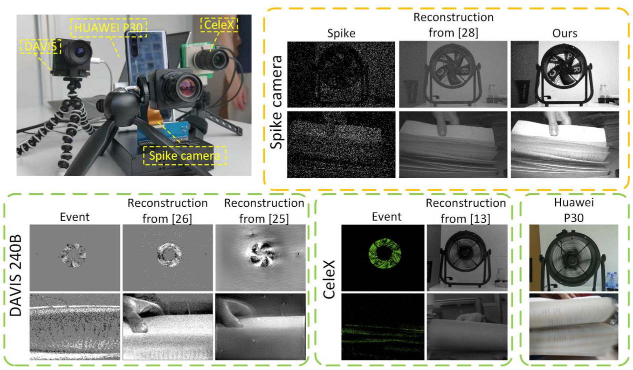

Figure 8. **The reconstruction results of different vision sensors.**

There are also some limitations in our method: some mechanisms in our model (e.g. motion confidence matrix and STDP) need a short period of previous spike data (about 15ms) to calculate the current state. If the camera itself moves too fast, the spike train received by the neuron may comes from different objects thus causing blur. Despite of this, we think it does not affect the application.

5. Conclusion

In this paper, we have proposed a novel three-layer spiking neural model to reconstruct visual images for spike camera. We comprehensively discuss the spike distribution and construct a probability model to describe it. Additionally, a dynamic neuron extraction model is proposed to distinguish the dynamic and static neurons. A combination of biologically plausible mechanisms in introduced to process the continuous-time spikes. Finally, the visual image can be reconstructed according to the state of the neuron and the firing threshold. To test our method, we build a dataset including normal-speed and high-speed scenes. The results show that our method can reconstruct high quality visual images in both high-speed motion and static scenes.

Acknowledgments. This work is partially supported by grants from the National Natural Science Foundation of China under contract No. 61825101 and No. U1611461.

References

[1] Patrick Lichtsteiner an Christoph Posch and Tobi Delbruck. A 128×128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, 2008. 1

[2] Juan Antonio Lenero Bardallo, Jose-Maria Guerrero-Rodriguez, Ricardo Carmona-Galan, and Angel Rodriguez-Vazquez. On the analysis and detection of flames with an asynchronous spiking image sensor. *IEEE Sensors Journal*, 18(16):6588–6595, Aug 2018. 1

[3] Michel Baudry. Synaptic plasticity and learning and memory: 15 years of progress. *Neurobiology of learning and memory*, 70(1-2):113–118, 1998. 3, 5

[4] Guo-qiang Bi and Mu-ming Poo. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, 18(24):10464–10472, 1998. 3, 4, 5

[5] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240×180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 1

[6] Eugenio Culurciello, Ralph Etienne-Cummings, and Kwabena A Boahen. A biomorphic digital image sensor. *IEEE Journal of Solid-State Circuits*, 38(2):281–294, 2003. 1

[7] F Yu Daniel and Jeffrey A Fessler. Mean and variance of single photon counting with deadtime. *Physics in Medicine & Biology*, 45(7):2043, 2000. 2

[8] Tobi Delbrück, Bernabe Linares-Barranco, Eugenio Culurciello, and Christoph Posch. Activity-driven, event-based vision sensors. In *IEEE International Symposium on Circuits and Systems*, pages 2426–2429, 2010. 1

[9] Peter U Diehl and Matthew Cook. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Frontiers in Computational Neuroscience*, 9:99, 2015. 3, 4, 5

[10] Pierre Falez, Pierre Tirilly, Ioan Marius Bilasco, Philippe Devienne, and Pierre Boulet. Mastering the output frequency in spiking neural networks. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018. 5

[11] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Asynchronous, photometric feature tracking using events and frames. In *2018 European Conference on Computer Vision (ECCV)*, pages 750–765, 2018. 1

[12] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984. 3, 4

[13] Menghan Guo, Jing Huang, and Shoushun Chen. Live demonstration: A $768\times$ 640 pixels 200meps dynamic vision sensor. In *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–1. IEEE, 2017. 1

[14] Christof Koch and Idan Segev. *Methods in neuronal modeling: from ions to networks*. MIT press, 1998. 3

[15] Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimizedvia graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):147–159, 2004. 5

[16] Juan Antonio Leero-Bardallo, D. H. Bryn, and Philipp Hfliger. Bio-inspired asynchronous pixel event tricolor vision sensor. *IEEE Transactions on Biomedical Circuits and Systems*, 8(3):345–357, June 2014. 1

[17] Stan Z Li. *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009. 4

[18] Li Xi, Liu Guosui, and Jinlin Ni. Autofocusing of isar images based on entropy minimization. *IEEE Transactions on Aerospace and Electronic Systems*, 35(4):1240–1252, Oct 1999. 7

[19] Martin Litzenberger, Christoph Posch, D Bauer, Ahmed Nabil Belbachir, P Schon, B Kohn, and H Garn. Embedded vision system for real-time object tracking using an asynchronous transient vision sensor. In *Digital Signal Processing Workshop-signal Processing Education Workshop*, pages 173–178, 2006. 1

[20] Lixiong Liu, Yi Hua, Qingjie Zhao, Hua Huang, and Alan Conrad Bovik. Blind image quality assessment by relative gradient statistics and adaboosting neural network. *Signal Processing: Image Communication*, 40:1–15, 2016. 7

[21] Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5419–5427, 2018. 1

[22] Niranjan D Narvekar and Lina J Karam. A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). *IEEE Transactions on Image Processing*, 20(9):2678–2683, 2011. 7, 8

[23] Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt. An asynchronous time-based image sensor. *IEEE International Symposium on Circuits and Systems*, pages 2130–2133, 2008. 1

[24] Stanislav Pyatykh, Jürgen Hesser, and Lei Zheng. Image noise level estimation by principal component analysis. *IEEE Transactions on Image Processing*, 22(2):687–699, 2012. 8

[25] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3857–3866, 2019. 8

[26] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *2018 Asian Conference on Computer Vision (ACCV)*, pages 308–324. Springer, 2018. 8

[27] Marcel Stimberg, Romain Brette, and Dan Goodman. Brian 2: an intuitive and efficient neural simulator. *BioRxiv*, page 595710, 2019. 7

[28] Lin Zhu, Siwei Dong, Tiejun Huang, and Yonghong Tian. A retina-inspired sampling method for visual texture reconstruction. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1432–1437, 2019. 1, 2, 7