Contents lists available at ScienceDirect

Engineering

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

Research
Artificial Intelligence—Article

基于普通器件实现快1000倍的相机与机器视觉

黄铁军，郑雅菁，余肇飞*，陈瑞，李源，李源，马雷，赵君伟，董思维，朱林，李家宁，贾杉杉，付溢华，付溢华，吴思，田永鸿

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

关键词
Vidar 相机
脉冲神经网络
超级视觉系统
全时成像

摘要

在数码相机中，我们发现了一个重大缺陷，即从胶片相机继承的图像和视频模型阻碍了相机捕捉快速变化的光子世界。我们提出了一种新的视觉形式，称为视象（vform），这是一个比特序列阵列，其中每个比特表示光子的累积是否达到了一个阈值，从而可以记录和重建任何时刻场景的光强。仅使用消费级CMOS（互补金属氧化物半导体器件）传感器和集成电路，开发了一种比传统相机快1000倍的脉冲相机。将视象看作生物视觉中的脉冲序列，进一步开发了基于脉冲神经网络的机器视觉系统，它可以将机器的速度和生物视觉的机理结合起来，从而实现了比人类视觉快1000倍的高速目标检测和跟踪，并通过辅助裁判和目标瞄准系统证明了脉冲相机和超级视觉系统的效用。视象模型和芯片有望从根本上改变图像和视频的概念以及摄影、电影和视觉媒体等相关行业，并开启一个全新的基于脉冲神经网络的速度自由的机器视觉时代。

1. 引言

数码相机是真正数码的吗？通常回答是肯定的，因为数码相机利用CCD（电荷耦合器件）/CMOS（互补金属氧化物半导体器件）传感器和数字电路成像取代了胶片上的成像。然而，数码相机本质上仍停留在模拟时代，因为它毫无保留地继承了图像和视频这种表达视觉信息的形式，这种形式对于胶片记录光的时间动态来说是必需的[1–3]，但对于纯数码系统并不必要。事实上，一副图像无法记录曝光时间内光线的变化，一段视频甚至会丢失相邻两次曝光之间的所有动态信息。此外，帧率只有几十赫兹的相机是无法拍摄高速场景的。与之相比，高速相机的时间采样频率可以达到数千赫兹甚至数万赫兹，但它们需要专用的传感器和快门，因此成本很高[4–5]。由此可见，图像和视频形式已经成为数码相机捕捉快速变化的光子世界的最大障碍。

我们提出了一种全新的视觉形式，称为视象（vform）。它打破了传统的基于帧的表示方法，使得我们可以研制成本效益高的高速相机。视象受灵长类视网膜中心凹的采样机制的启发[6–7]，采用脉冲序列表达光的变化过程，能够有效保留物理光流的时序和更为准确的时间信息，可以从中构建出任何时刻的图像，我们称这种能力为全时成像（fulltime imaging）。

基于视象模型，我们采用传统相机同样的CMOS传

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

英文原文：Engineering 2023, 25(6): 110–119
引用本文：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. 1000× Faster Camera and Machine Vision with Ordinary Devices. *Engineering*, https://doi.org/10.1016/j.eng.2022.01.012

感器和消费级的集成电路[8]，研制出了VidarOne芯片和脉冲相机。当光敏器件累积收集的强度超过约定阈值时就产生一个脉冲。普通光敏器件的光电转换速度在10 ns量级，相比之下，人类视网膜完成光电转换需要10 ms，这里存在6个数量级的差异[9]，因此虽然有类似的机制，脉冲相机避免了生物视觉的速度限制。我们研制的首款脉冲相机的时间采样频率为40 000 Hz，实现了超过人类视觉三个数量级的高速成像。

脉冲相机产生的脉冲流具有清晰的物理意义，它对输入场景的时空视觉信息进行编码，因此可以用来执行高速视觉任务。然而，基于人工神经网络（ANN）的传统机器视觉方法需要将脉冲流转换为图像（40 000帧 · s^{-1}），然后逐帧处理，因此不能实时处理这些脉冲流[10]。我们发现脉冲神经网络（SNN）可以自然地实时处理脉冲相机的输出脉冲流[11–12]。利用这种方法，我们开发了一个基于SNN的超级视觉系统，它可以将机器的速度和生物视觉的机理结合起来[13–18]。所谓视觉，就是脉冲序列在这种网络上的流转过程，因此速度只取决于神经网络的物理性能。目前我们在普通CPU模拟的脉冲神经网络上已经实现了对上述40 000 Hz脉冲流实时处理，实现了速度超越人类视觉三个数量级的高速视觉。未来采用SNN硬件和更高速的脉冲相机，可以实现人类视觉更多数量级的超级视觉，而所需要的只是目前已经广泛使用的消费电子级别的光电器件和电路技术。

2. 方法

（1）视觉内容重构。TFW (texture from window)方法通过计算一个时间窗口中的脉冲数目来获得像素值（与场景的光强成比例）。具体来说，移动时间窗收集特定时段的脉冲。通过计算这些脉冲，像素值可通过下式来估计：

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

式中，P_{t_i}表示t_i时刻的像素值；w表示时间窗的大小；N_w表示时间窗内收集的所有的脉冲；C表示重建图像的最大动态范围。TFI方法假设场景光强$\bar{I}$在一个很短时间内是一个常数。根据脉冲相机的机理，脉冲发放条件可以简化为$\bar{I}\Delta t \geqslant \phi$，这里$\Delta t$表示两个脉冲之间的时间间隔，$\phi$表示阈值，因此可以通过两个脉冲来估计像素值：

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

式中，Δt_i表示t_i时刻对应的两个脉冲之间的时间间隔。

我们对提出的图像重建算法进行了测试，并与传统相机进行了比较。我们构建了一个混合相机系统，由脉冲相机、传统相机和分光镜组成。通过分光镜，两个相机可以记录相同的场景。我们采用了两种无参考图像质量评估指标，即二维（2D）熵和标准差（STD）。2D熵使用像素的灰度值和其局部平均灰度值来评估图像所携带的信息量，较大的2D熵表示更多的信息。STD评估图像的对比度，较大的STD表示较高的对比度。如表1所示，在这两个指标中，我们的重构方法在所有方面都取得了比传统相机更好的结果。

表1 TFI、TFW和传统相机之间的比较

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

（2）动态连接门。动态连接门是基于短期可塑性（short-term plasticity, STP）的一种结构。STP是指突触强度的短期（通常在几十到数千毫秒之间）变化，也被称为神经元之间的动态连接[19–20]。当突触后神经元从突触前神经元接收到一系列动作电位时，突触后电位（post-synaptic potential, PSP）会根据以下方程发生变化：

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

式中，A是动作电位可在突触后神经元上触发的最大电流值；$x(t)$ [$0 < x(t) < 1$]表示t时刻轴突末端中可用神经递质的剩余数量；$u(t)$表示t时刻轴突中神经递质的释放概率。当突触后神经元从突触前神经元接收到具有固定频率的动作电位序列时，PSP在几个脉冲到达之后会收敛到稳定状态[21] [图1（a）]。如果突触前神经元的脉冲发放频率发生变化，则PSP将在稳定值附近波动[图1（b）]。通过利用STP对输入脉冲流的发放时间模式的敏感性，可以对由背景或静态区域生成的脉冲流进行过滤，仅保留由运动对象生成的脉冲流。

（3）检测与跟踪。滤波层中的神经元连接到检测层中的9个相邻的LIF神经元。每个LIF神经元会累计从突触前神经元传来的电流，并在膜电位达到阈值时触发脉冲。由于只有运动物体相对应的区域会生成脉冲流，因此可以通过检测发放神经元的连接区域来找到每个移动物体。最后通过基于检测所得结果来跟踪不同的运动对象。为了评估算法的准确性，我们使用检测成功率（DSR）来衡量物体检测的效果，并使用多物体跟踪准确率（MOTA）、误报率（FP）、漏检率（FN）和标识符切换率（IDS）来评估物体跟踪的效果。结果如表2所示，我们可以发现，我

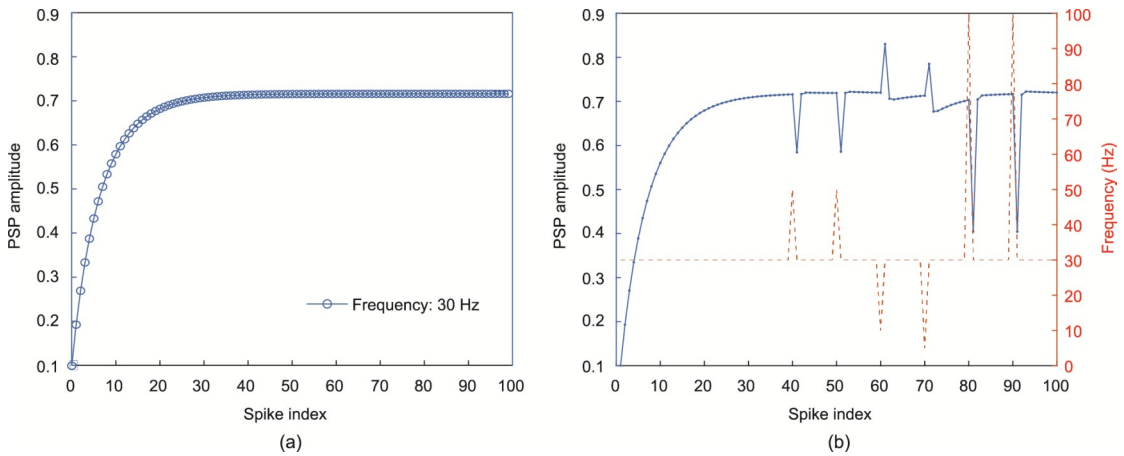


图1. PSP根据前突触神经元传来的脉冲变化。（a）当输入脉冲的频率为固定值30 Hz时，PSP（蓝色曲线）收敛到一个稳定值；（b）当输入脉冲的频率发生变化时（红色曲线），PSP（蓝色曲线）在稳定值附近震荡。

们的算法可以在低功耗的情况下实现良好的性能。

表2 检测与跟踪的准确性。检测与跟踪的定量评估

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

（4）用于预测的连续吸引神经网络。连续吸引神经网络（CANN）是用于神经信息表示的一种经典网络模型。先前的研究表明，通过在神经元动力学中添加负反馈，CANN可以以近似恒定的领先时间来预期地跟踪运动物体[22]。在此基础上，我们提出了一种改进的CANN模型，该模型的视觉输入直接来自于脉冲相机，可用于预测跟踪快速移动的对象。

（5）目标识别。识别脉冲神经网络的突触权值是用BP-STDP进行训练的，该规则来自Tavanaei和Maidi[23]。在这里，我们在最后一层中使用多个脉冲神经元来表示一个类别。具体来说，m个神经元被分成n个组来表示n个类别（$m=kn$）。我们利用脉冲相机生成的真实的脉冲数据与Spike-Sim生成的模拟数据作为训练数据集。在训练过程中，目标组中膜电位最大的分类神经元根据STDP更新其突触权值[24]，而非目标组中膜电位最大的不发放神经元则根据anti-STDP更新其突触权值[25]。然后根据前突触活动，将突触权重的调节逐层反向传播。

（6）估计速度。由于风扇的角速度为每分钟2400转，并且字符中心到风扇中心的距离为0.12 m，因此可以估计风扇上字符的线速度为$2400/60\times2\times\pi\times0.12\approx30\,(\mathrm{m\cdot s^{-1}})$。考虑到风扇与脉冲相机之间的距离为0.75 m，基于中心透视原理，脉冲相机和超级视觉系统可以在1 m范围内实时检测、跟踪和识别线速度为40 $\mathrm{m\cdot s^{-1}}$的运动目标。

（7）Spike camera high-speed spike dataset (SCHSSD)。该数据集包括：①使用静止脉冲相机（Class A）捕获的高速运动目标的脉冲流；②使用高速运动脉冲相机（Class B）捕获的自然场景的脉冲流。Class A包含一辆行驶的汽车、一个旋转的圆盘、一个旋转的风扇和一个爆破气球，Class B包含火车、森林、高架桥和铁路场景（更多细节见表3）。我们还提供SpikePlayer软件来播放脉冲序列。

表3 VHSSD数据集描述

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

（8）SpikePlayer。这是一个可视化软件，它可以播放真实和模拟的时空脉冲流（即.dat文件），可以提供基于TFW和TFI方法重构的高帧率视频。SpikePlayer支持不同的分辨率，如400×250，扩展了模拟器的兼容性。

（9）Spike-Sim。Spike-Sim是脉冲相机的模拟器，它可用于模拟3D场景中的任意相机运动和物体运动，同时提供参考图像和相机姿态、物体速度等附加信息。它集成了脉冲相机原理和多种渲染引擎，包括基于OpenGL开发的可以实时渲染和生成脉冲流的渲染器以及基于Blender循环引擎的照片级的真实感渲染。

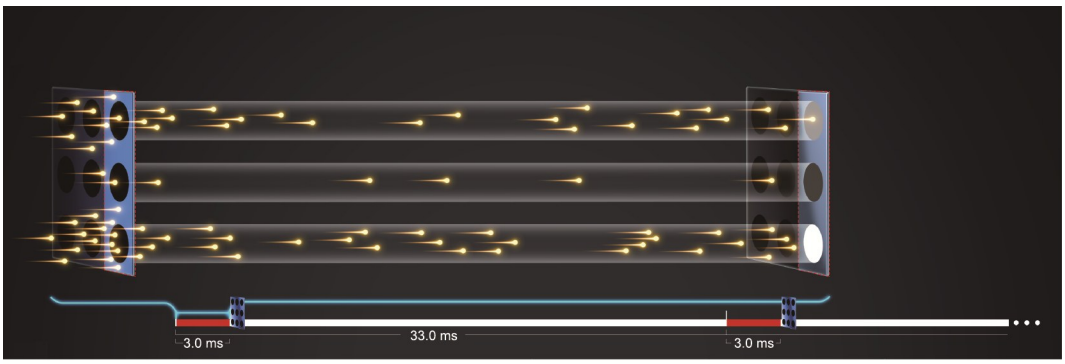
3. 结果

3.1. 视象——一种新的更自然的视觉形式

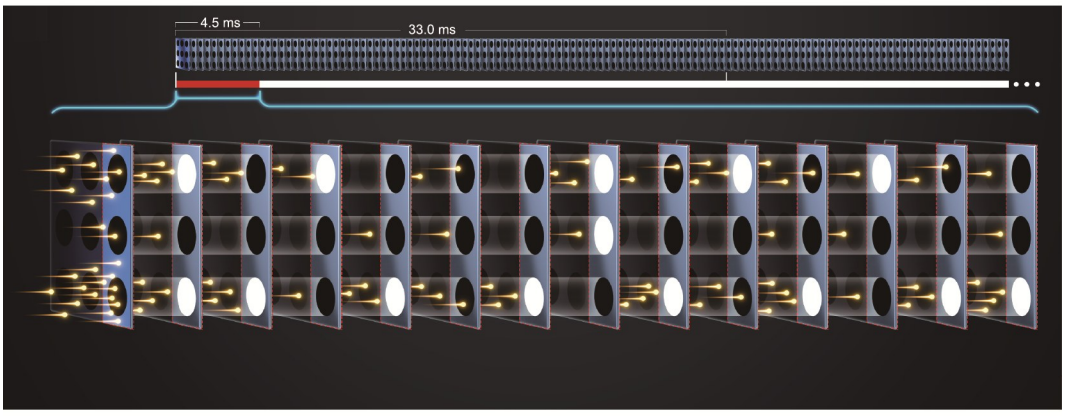
在我们介绍新的视觉形式视象之前，我们简要回顾图像和视频的概念。对于相机视锥范围内高速射入的大量光子[图2（a）]，相机按照预定帧率f，分别在时刻$t_1, t_2, t_3, \ldots$（时间间隔均为$1/f$ s）进行图像采集。图像采集时所有感光单元同时进行持续时长为Δt（即曝光时间，$\Delta t < 1/f$）的光子捕获，然后分别将累积的光强记录下来，按照感光单元空间排布而形成的光强分布就是图像，图像按等时间间隔排列成的序列就是视频。从全光函数的角度看[26]，图像记录的不是$t_1, t_2, t_3, \ldots$时刻的状态，而是持续Δt的物理过程的累积。对于视频来说，采集每帧图像所用的累积时间Δt小于等于两帧图像之间的时间间隔$1/f$，这意味着，$1/f - \Delta t$这段时间内的信息彻底丢失了，Δt时段内的运动过程也被“压扁”到图像上而丢失了，因此视频的时域采样率不是对物理过程的一个完整采样。

图像和视频的同步曝光且曝光时间相同的设计已经成为数码相机捕捉快速变化的光子世界的最大障碍，这样的设计是没有必要的。这里我们介绍一种新的视觉表示方式，即视象。它使用新的时域采样机制并允许异步曝光，从而可以更好地捕捉光的时域变化。在这里，视象作为视觉和form的组合被创造出来，定义一种新的视觉信息形式来取代视频。

视象在空域采样方面与传统图像和视频并无二致，实际上，脉冲相机采用和普通相机完全相同的感光器件，也就是大家熟知的CMOS和CCD，因此视象也采用空间排布的点阵表达空域信息[图2（b）]。视象和视频的根本不同是采用新的时域采样方式，所有感光器件并不同步按相同的曝光时长进行曝光，而是持续捕获光子，当收集的光强超过约定阈值就产生一个脉冲[图2（b）]，这个脉冲及形成这个脉冲所持续的时长称为视元。每个感光器件产生的视元按照时间次序排成序列，最简形式就是一个比特流，1表示在该时刻出现了一个脉冲，0表示处于累积状



(a)



(b)

图2. 图像与视象在视觉信息表示方面的整体比较。（a）图像和视频的视觉表达。感光单元（三个圆）捕获一组光子（黄色流星）并在3 ms（红线）的曝光时间内输出累积强度（由圆的亮度表示）。图像就是按照感光单元空间排布而形成的光强分布，根据预先定义的帧率f=30 Hz，每33 ms采集图像得到的就是视频。需要注意的是30 ms这段时间内（白线）的信息彻底丢失了。（b）视象的视觉表达。感光单元（三个圆）持续捕获光子，当累积强度超过给定阈值（此处阈值为四个光子）时产生脉冲（白色圆）。视象是按照器件的空间排布组成的比特流阵列，其中比特1表示此时感光单元产生了一个脉冲，比特0表示感光单元尚处于累积状态。图中三个感光单元的视象可以表示为$\begin{bmatrix} 01010001010100 \\ 00000001000000 \\ 01101010101011 \end{bmatrix}$。

态，比特1和它之前另一个1之间的所有0构成一个数码视元。每个感光器件产生一个脉冲流，所有感光器件产生的脉冲流按照器件的空间排布组成脉冲流阵列，就是视象。

视象比视频突出的优势在于有效保留了各个采样位置光的时域变化过程。采用单光子敏感器件，一个光子就能激发一个脉冲，这时脉冲相机记录的就是精确完整的物理事实。普通感光器件捕获一组光子才能激发一个脉冲，是对物理事实的一种粗糙表示，但物理过程的时间关系仍然得到了最大程度的保留，而不像视频那样通过人为规定将时间关系整齐划一地强制为数十赫兹。事实上，今天广泛普及的CMOS感光器件，时间灵敏度已经达到数十纳秒，采用视象这种新模型，就可以实现千万赫兹的高速时域采样，记录极快的物理过程。当然，日常视觉应用并不需要这么高的采样频率，我们开发的第一款芯片设定的采样频率是40 000 Hz，也就是比人类视觉和普通相机快1000倍，已经可以清晰拍摄时速350 km的高铁和转速达每分钟7200转的硬盘。

视象记录了一定空间范围内各个位置光的精细变化过程，其物理意义十分清晰明确，因此，用视象生成传统的图像和视频就是意料之中的事情。事实上，对于任意指定时刻，可以从覆盖该时刻的视元估计该位置的相对光强，更可以参考之前的与空间上邻近的更多视元估计出更精细的光强，从而得到任意时刻的精细图像。视象这种蕴含了任意时刻影像的能力我们称为全时成像（fulltime imaging），或者说连续成像（continuous imaging）。

3.2. VidarOne芯片和脉冲相机系统

VidarOne芯片基于新的视觉表示模型视象开发，它采用了异步像素触发结构。如图3（a）所示，400×250像素阵列将输入光子转换成脉冲流阵列并且利用滚动快门来检测所有像素的响应。之后行扫描器逐行扫描像素阵列。当逻辑控制信号选择一行像素时，数据被传送到数字缓冲器中便于并行读出。为了支持脉冲流的高速输出，VidarOne芯片提供了一个8通道专用通信接口，带宽为每秒500兆比特（Mb）。同步读出接口的时钟频率为20 MHz。

一个像素的基本电路如图3（b）所示，它包括脉冲触发电路、复位电路和读出电路。像素电路中的光电二极管连续捕获光子并将入射光转换为连续的光电流I_{ph}。在这一过程中光电二极管的电压V_{pix}会持续下降。当光电二极管的电压V_{pix}达到给定阈值V_{ref}时，比较器的输出切换，并且产生一个翻转（脉冲）信号[图3（c）]。锁存器在时钟信号*clk*的使能操作下同步比较器的翻转信号。一旦锁存器检测到翻转信号，光电二极管的电压V_{pix}被重置为预先定义的重置电压。同时，脉冲信号被发送到RS触发器并被保存。行读出信号R_d控制脉冲流的顺序扫描和读出，行复位信号R_{st}负责清除RS触发器中的信号。脉冲像素的时序图如图3（d）所示，在时钟信号clk的控制下，脉冲信号被固定在持续100 ns的高电平。对于在时刻A产生的脉冲信号（箭头A），考虑到行复位信号R_{st}处于低电平，脉冲在50 ns之后由高电平的行读出信号R_d读出。对于在时刻B产生的脉冲信号（箭头B），考虑到行复位信号R_{st}处于高电平，RS触发器被屏蔽，所以RS触发器会在50 ns后捕获脉冲信号。当下一个行读出信号R_d到达时，脉冲被读出。如果一个信号周期中有两个或两个以上的脉冲信号，只有一个脉冲信号可以被处理。其原因是RS触发器在被脉冲信号锁存时不会对其他脉冲做出响应。由于行扫描时间为100 ns，250行像素阵列产生的脉冲流的时间分辨率为25 μs。

我们采用标准的110 nm 1-poly 3-metal制造工艺，VidarOne芯片的芯片面积为9.96 mm×7.1 mm [图3（e）]。每个正方形像素的尺寸为20 μm×20 μm，在原型芯片上实现了13.75%的填充因子。大尺寸的像素探测器可以保证放置金属网格后有足够的光电探测器面积。在自然光条件下，该芯片不需要任何动态范围增强技术就可以提供超过100 dB的高动态范围成像。拟定设计的能耗约为370 mW。封装的VidarOne芯片的物理视图如图3（f）所示。布局和布线经过精心设计，使得像素电路的硅面积最小。

我们设计研制了具备高速摄像功能的轻量化脉冲相机系统，它主要由视觉信息采集模块、高速传感模块和实时视觉计算模块组成[图3（g）]。视觉信息采集模块将输入场景转换为脉冲流，然后未经加工的高速脉冲序列经过传感模块进行高吞吐率实时数据处理操作，并通过PCIE总线接口发送给视觉计算模块。

3.3. 脉冲相机视觉场景重构

脉冲相机具有全时成像的能力，可以根据输出脉冲流（vits）的特性来重构任意时刻的视觉场景，重构场景的动态范围和质量都非常灵活。为了重建脉冲相机拍摄的场景，弥补视象数据与传统的基于帧的视觉之间的差异，我们提出了两种视觉场景重构方法：基于窗口的场景重构（texture from window, TFW）[图4（a）]和基于脉冲间隔的场景重构（texture from interspike interval, TFI）[图4（b）]。更多细节见第2节。

具体来说，TFW方法利用了场景光强与脉冲计数（发放率）成正比的原理，因此我们可以利用移动的时间窗口来收集特定周期内的脉冲数量来反推像素值（与光强

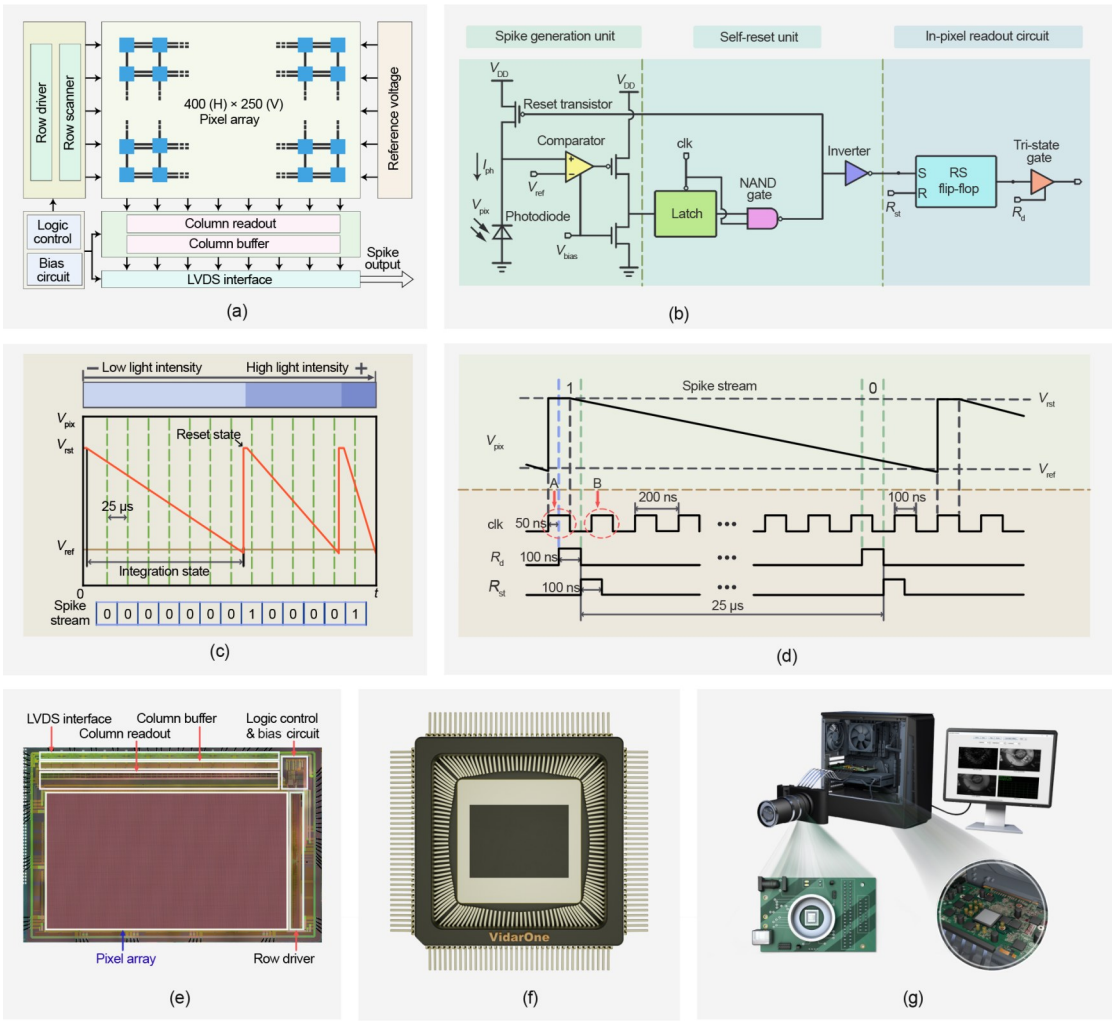


图3. 设计VidarOne芯片和脉冲相机系统。（a）芯片架构示意图。它主要由像素阵列、具有配置驱动器的行扫描器、具有可寻址数字缓冲器的列读出电路、偏置/基准电路和数字逻辑控制电路组成。（b）像素电路包括三部分：由光电二极管、复位晶体管和比较器组成的脉冲触发电路；由锁存器、与非门和反相器组成的复位电路；由RS触发器和三态门组成的读出电路。V_{pix}和V_{ref}分别表示光电二极管的电压和参考（阈值）电压；V_{DD}和V_{bias}分别是电源电压和偏置电压；I_{ph}是光电流；C_r和C_d是用于抑制电压波动的电容器；R_{st}是行复位信号；R_d是行读出信号；clk表示时钟信号，当clk处于高电平时，锁存器将被锁定。（c）脉冲触发和脉冲编码原理。V_{rst}是重置状态。像素强度在触发翻转信号时被编码为1，否则被编码为0。（d）脉冲流的时序图。（e）VidarOne芯片的显微照片。（f）封装好的VidarOne芯片的图像。（g）脉冲相机系统包括由工业摄像机镜头和VidarOne芯片组成的视觉信息采集模块以及由FPGA芯片实现的高速传感模块和由台式机工作站实现的实时视觉计算模块。

成正比）[图4（a）]。重构结果如图4（c）所示，这里我们给出了一个新的脉冲相机数据集并称其为脉冲相机高速数据集（spike camera high-speed dataset, SCHSSD）（见第2节）。图4（c）的第一行展示了脉冲相机拍摄8个不同场景产生的原始数据，第二行展示了利用TFW方法重构的场景。TFW方法适用于静态场景。在高速运动场景中，脉冲相机接收到的光强变化很快，此时一段时间内的平均发放率不能很好地捕捉场景光强的快速变化，从而会引起成像模糊[图4（c）的第二行]。TFI方法通过利用场景光强与脉冲间隔成反比的原理解决了这个问题[图4（b）]。因此，我们只通过两个脉冲（即一个脉冲间隔）就可以估计出场景某一时刻的光强。该方法也与高速运动场景中场景照明的快速变化相匹配。对于高速运动场景，TFI方法可以获得比TFW方法更好的结果[图4（c）的第三行]。我们还将我们的构建结果与传统相机的构建结果进行了定量比较。如表1所示，我们的重构方法可以获得比传统相机更好的结果。

为了便于验证新的想法，我们开发了一个脉冲相机模拟器Spike-Sim，它可以模拟3D场景中任意的相机运动和物体运动，并生成类似于脉冲相机拍摄的可靠的脉冲序列（见第2节）。此外，它通过模拟像素的RGB通道来提供彩色图像。在这里我们用Blender构建了“北大飞球”和“北大硬币”的场景，并用Spike-Sim生成了脉冲序列。图4（d）的第一行和第二行分别呈现由普通照相机（帧率为f=30 Hz）生成的参考图像和由Spike-Sim生成的模拟脉冲序列。基于TFW方法和TFI方法重构的结果展示

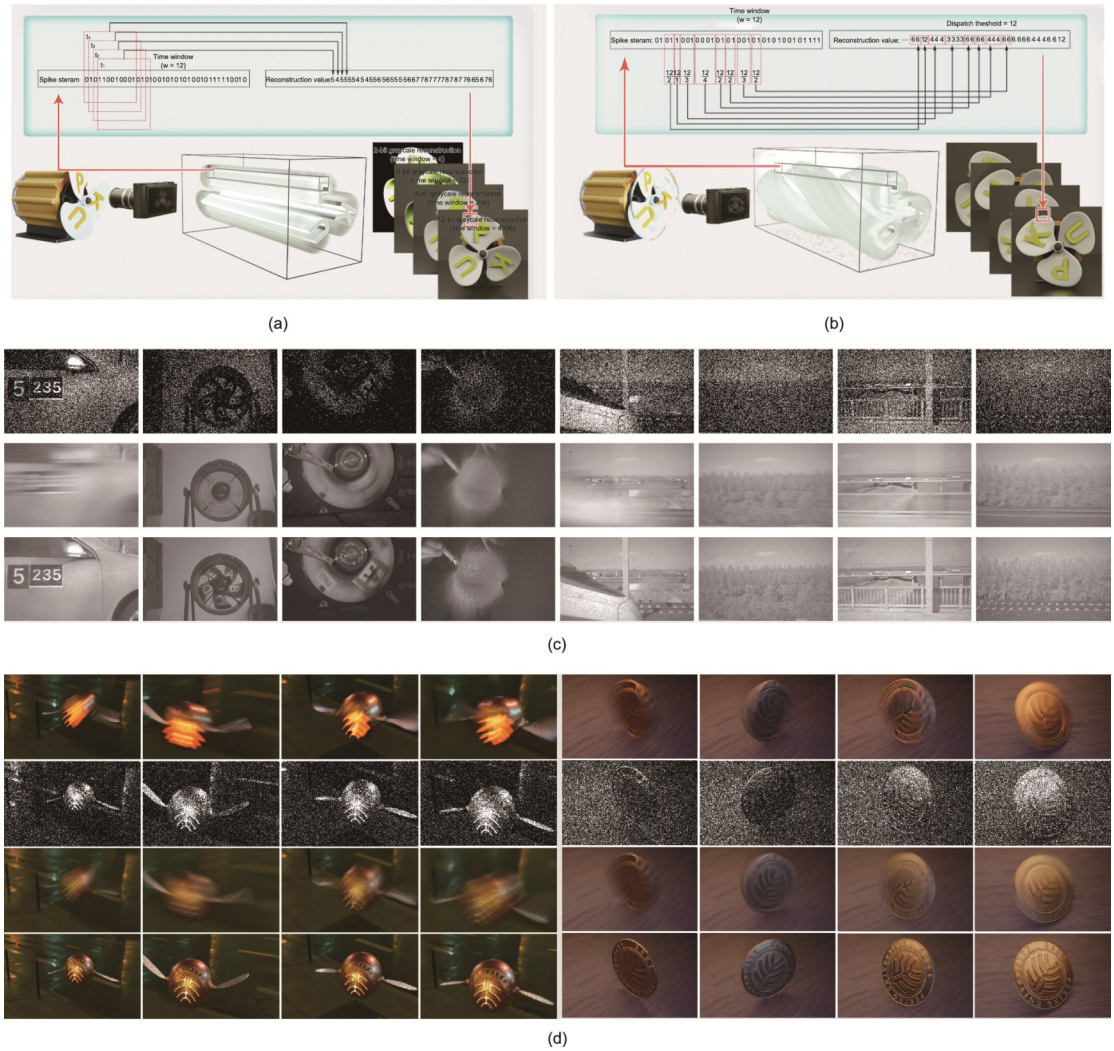


图4. 脉冲相机视觉场景重构。（a）TFW方法原理图。它基于场景光强与脉冲数量成正比。浅蓝色矩形表示其中一个像素产生的脉冲流和相应的重建的灰度值。TFW方法可以通过调整时间窗口的宽度来收集不同数量的脉冲（请参见右侧的四帧），从而重构具有自由动态范围的场景。（b）TFI方法原理图。它基于场景光强与脉冲间隔成反比，适用于高速运动场景。（c）VHSSD数据集重构结果。三行分别表示脉冲相机产生的原始数据、基于TFW方法重构的场景以及基于TFI方法重构的场景。（d）利用Blender构造的两个场景的重建结果。场景一：“北大飞球”（灵感来源于《哈利波特》系列中的“金色飞贼”魁地奇游戏球）。飞行的球在夜间拍打翅膀，从远到近进入脉冲相机的视野。场景二：“北大投币”。一枚印有北大标志的金币在木制桌面上旋转，最后停了下来。两个场景都考虑了脉冲相机的轻微移动。

在第三行和第四行中。对比可见参考图像中的细节是模糊的，而基于Spike-Sim生成的脉冲流重建的图像可以显示更多的纹理细节。

3.4. 基于SNN的超级视觉系统

本节我们展示如何通过结合机器的速度和生物视觉系统的机制来实现超级视觉。我们提出了一种基于脉冲神经网络的超级视觉系统，可用于高速运动物体检测、跟踪、预测和识别，该系统比人类视觉快1000倍[图5（a）]。实现这些功能主要包括三个挑战：首先，消除从场景的背景/静态部分产生的脉冲，以便进行后续的高级视觉任务；其次，检测并平滑跟踪高速运动的物体并预测其轨迹；再次，识别被跟踪的物体。为了完成这些任务，我们提出了一种基于短期可塑性的动态连接门并用于过滤时空脉冲序列；提出了用于对象检测和跟踪的局部连接的脉冲神经网络模型；提出了用于预测运动的连续吸引神经网络；还提出了用于目标识别的三层全连接脉冲神经网络模型。

超级视觉系统的详细结构如图5（b）~（e）所示。对于脉冲相机，场景的背景/静态部分也会输出具有固定频率的脉冲序列，这会妨碍后续高级视觉任务的执行。因此，我们在此处引入基于短期可塑性的动态连接门来过滤脉冲序列[图5（b），细节见第2节]。当输入脉冲流具有固定频率时（对应于背景或静态对象），门将关闭；当脉冲流频率变化时（对应于运动对象），门将打开。因此动态连接门仅保留了由运动部分产生的脉冲流。滤波层中的神经元将兴奋性突触后电位（EPSP）传输到检测层中相邻的神经

元，所有神经元根据漏电积分发放（leaky integrate-and-fire, LIF）模型产生脉冲[图5（c）]，通过检测发放的神经元对应的区域可以找到每个运动物体。在跟踪层中，可以通过比较先前时间与当前时间的运动神经元的位置或拓扑相似性来关联不同的运动物体。下一层是连续吸引子网络CANN [图5（d）]，它通过添加负反馈来预测轨迹。CANN可以以近似恒定的领先时间来预期地跟踪运动物体（见第2节）。识别网络是多层全连接的脉冲神经网络[图5（e）]，该网络使用BP-STDP学习规则进行训练（见第2节），并根据网络最后一层神经元的脉冲发放率确定识别结果。

3.5. 脉冲相机和超级视觉系统的效用验证

我们通过辅助裁判和目标瞄准系统来证明脉冲相机和超级视觉系统的效用。图6（a）显示了辅助裁判场景，在这个场景中，我们使用乒乓球发球机发射一个球来模拟网球和羽毛球等球类运动。我们研究的问题是确定球落地时是在界内还是界外（白线）。当球的着地位置接近界线时，人眼很难做出判断，因此在比赛中经常使用鹰眼系统。事实上鹰眼系统无法记录球着地的瞬间，一般是根据运动轨迹来估算着地位置，所以有时引起纠纷，此外，鹰眼系统价格非常昂贵。相比之下，脉冲相机具有全时成像能力，能够记录球下降的整个过程，使裁判员能够确定球着地的位置[图6（b）]。

此外，我们搭建了目标瞄准系统来证明脉冲相机和超级视觉系统的组合可以实现高速视觉[图6（c）]。脉冲相机放置在高速旋转（转速约2400 $r \cdot min^{-1}$）的风扇前面，其叶片上粘贴有三个字符（“P”“K”和“U”）。可以事先选取其中一个字符做为瞄准目标，激光器需要发射激光脉冲并击中该字符上面的相纸。解决这一问题需要三个步骤：首先，检测并跟踪场景中所有的运动物体；其次，识别所有的运动物体并确定预先给定的字符的位置；再次，预测字符的运动轨迹并控制激光击中目标。我们利用超级视觉系统（图5）来完成该任务。图6（d）显示了输出层的神经元对于不同的字符产生的响应脉冲。图6（e）强调了三个字符的检测和跟踪性能，从中可以发现网络可以检测所有运动目标并平滑地跟踪它们。图6（f）给出了激光击中前后风扇的比较。该系统为评估脉冲相机和超级视觉系统的性能提供了一种合适的方法。脉冲相机和超级视觉系统可以实时检测、跟踪和识别在0.75 m范围内以

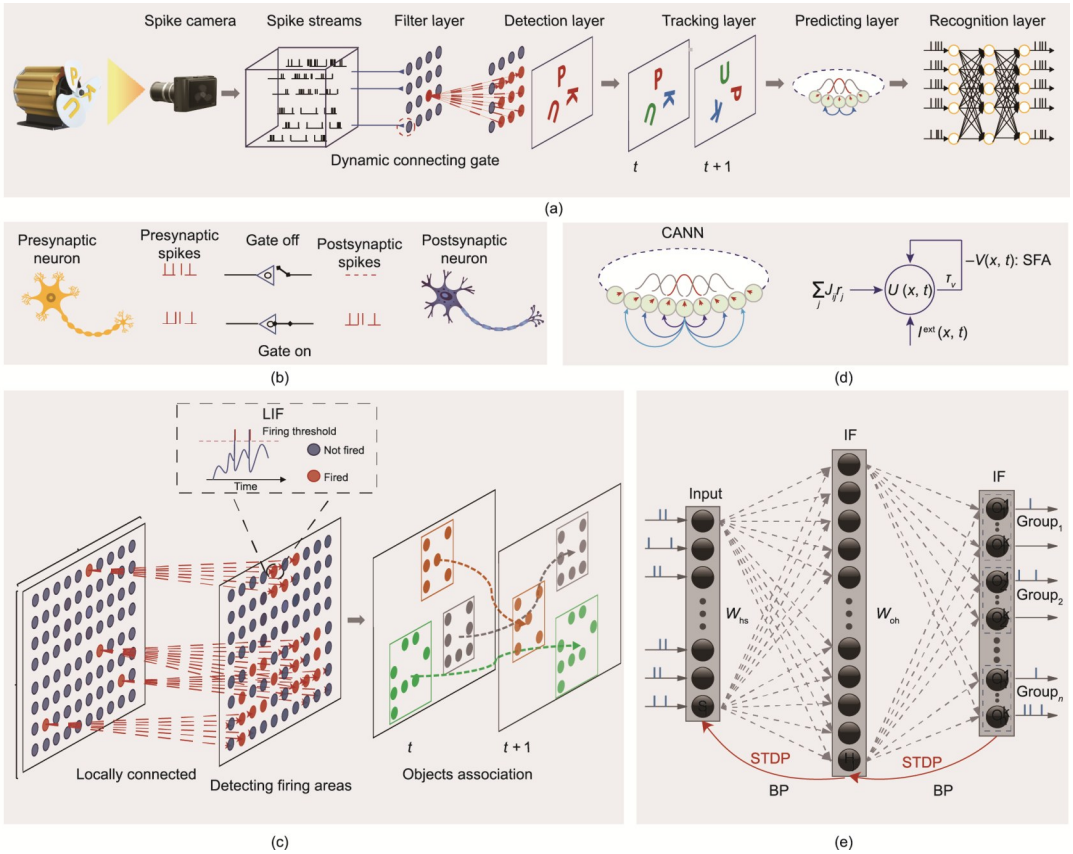


图5. 超级视觉系统。（a）基于SNN的高速运动物体检测、跟踪、预测和识别的框架；（b）基于短时可塑性的动态连接门可用于过滤时空脉冲序列，它可以过滤具有固定发放频率的脉冲流；（c）用于预测轨迹的CANN；（d）用于物体检测跟踪的局部连接脉冲神经网络；（e）用于物体识别的三层全连接脉冲神经网络。

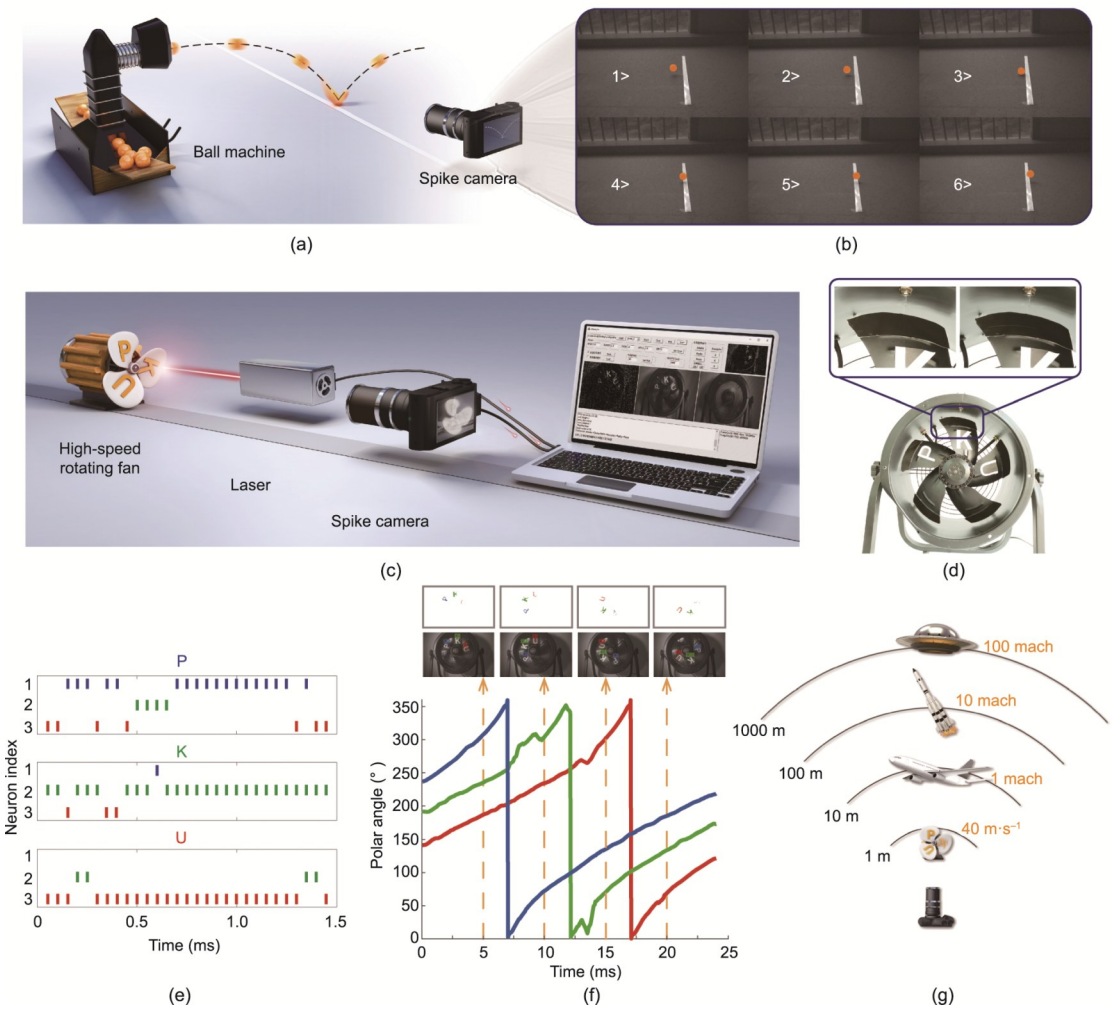


图 6. 通过辅助裁判和目标瞄准系统来证明脉冲相机和超级视觉系统的效用。（a）辅助裁判任务，脉冲相机用于检测球落地时是在界内还是界外。（b）脉冲相机可以记录球下降的整个过程，这里乒乓球的速度约为 100 km·h^{-1}。球和边界是彩色的是用于强调，我们仅展示了 170 帧里面的 6 帧。（c）激光击中前后风扇的比较。激光发射了 64 个脉冲，都击中了事先指定的字符“K”。（d）目标瞄准系统。激光需要击中高速旋转风扇上指定字符旁边的相纸。（e）脉冲神经网络识别测试。脉冲神经网络输出层对应正确类别的神经元发放的脉冲最多。（f）多目标检测跟踪。y 轴表示每个目标中心点相对于风扇中心的极角。脉冲神经网络可以获得每个字符的掩膜并实时返回其边界框。掩膜和边界框的不同颜色对应于不同的物体。（g）评估脉冲相机和超级视觉系统的性能。

30 m·s^{-1} 的线速度转动的风扇（见第 2 节）。根据中心透视原理，它能对 10 m 以内的以声速飞行的飞机进行实时检测、跟踪和识别，并能对 1 km 以内的以 100 马赫高速运动的目标进行实时检测、跟踪和识别[图 6（g）]。

3.6. 应用前景

脉冲相机机能够捕捉快速的物体运动。它具有与人眼功能相似的高速模式，性能优于人眼。传统的基于帧的相机是难以企及的，因为帧与帧之间会丢失大量的信息。通过提高帧速率，一些高速摄影机（如幻影摄影机）缓解了这一问题，但它们需要专用的传感器和快门，而这些传感器和快门往往非常昂贵。相比之下，脉冲相机是采用传统的 CCD 传感器和常规的半导体制造工艺，更具成本效益。因此，它可以广泛应用于日常生活中，如手机和相机。

相比于传统相机，脉冲相机的另一个优点是它提供了更灵活的图像采集方法。脉冲相机可以在任意时刻重建图像，其动态范围有相当大的灵活性。动态视觉传感器（dynamic vision sensor, DVS）是另外一种受视网膜启发的相机[20–22]，它的感光单元仅在亮度变化超过某个阈值时才会产生事件。区别于此，脉冲相机的每个感光单元不断捕获光子，当累积强度超过给定阈值时产生脉冲。因此，脉冲相机可以有效地记录每个采样位置的光强。我们相信，脉冲相机将在监控系统领域创造新的生态，可以应用于动态人脸识别、指纹识别和掌纹识别。

脉冲相机是受灵长类中央凹的神经回路结构和信息处理机制启发而提出的，它可以将光信号转换成电信号并输

出脉冲序列。这些脉冲序列可以自然地被脉冲神经网络处理。鉴于脉冲神经网络在视觉感知和认知任务中具有高效性和有效性，我们期望脉冲相机和基于脉冲神经网络的超级视觉系统的结合将为基础研究问题和实际应用提供丰富的实用工具，如以电子速度实现目标检测、跟踪和识别。

4. 结论

通过用视象代替视频，脉冲相机将相机发展拉回正确轨道，将释放光电技术被压抑数十年的技术潜力，在几乎所有领域替代传统视频相机，引发相机领域的一场革命。

视象本质上是表征光学时空变化过程的脉冲序列，正好作为脉冲神经网络的输入，脉冲相机才是机器视觉的眼睛，将在人工智能时代发挥重要作用。

致谢

本工作得到了国家自然科学基金项目（61425025）和北京市科技计划项目（Z151100000915070和Z1711000001170 08）的支持。

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 theproctor lecture. InvestOphthalmol 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 spiketiming- 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.