Journal of System Simulation

Volume 36 | Issue 3 Article 10

3-15-2024

Gesture Recognition for Dynamic Vision Sensor Based on Multi-dimensional Projection Spatiotemporal Event Frame

Lai Kang
College of Systems Engineering, National University of Defense Technology, Changsha 410073, China; Laboratory for Big Data and Decision, National University of Defense Technology, Changsha 410073, China, kanglai@nudt.edu.cn

Yakun Zhang
PLA 61081 Troops, Beijing 100089, China

Gesture Recognition for Dynamic Vision Sensor Based on Multi-dimensional Projection Spatiotemporal Event Frame

Abstract

Abstract: Vision-based gesture recognition is a commonly used means of human-computer interaction in the fields of virtual reality and game simulation. In practical applications, rapid changes in gesture movements will lead to blurred imaging with traditional RGB cameras or depth cameras, which brings great challenges to gesture recognition. To solve the above problems, a dynamic visual data gesture recognition method based on a multi-dimensional projection spatiotemporal event frame (STEF) is proposed by a using dynamic vision sensor to capture high-speed gesture movement information. The spatiotemporal information is embedded in the data projection surface and fused to form a multidimensional projection STEF, which overcomes the limitation of the time-domain information loss of the existing event frame expression method of dynamic visual information and improves the feature expression ability of dynamic visual sensing data. On this basis, advanced spiking neural networks are used to classify STEFs to realize gesture recognition. The recognition accuracy of the above method on the public dataset reaches 96.67%, which is better than similar methods, indicating that the proposed method can significantly improve the accuracy of gesture recognition in dynamic visual sensing data.

Keywords

dynamic vision sensor, gesture recognition, multi-dimensional projection, spatiotemporal event frame, spiking neural network

Recommended Citation

Kang Lai, Zhang Yakun. Gesture Recognition for Dynamic Vision Sensor Based on Multi-dimensional Projection Spatiotemporal Event Frame[J]. Journal of System Simulation, 2024, 36(3): 649-658.

第 36 卷第 3 期 系统仿真学报© Vol. 36 No. 3
2024 年 3 月 Journal of System Simulation Mar. 2024

基于多维投影时空事件帧的动态视觉传感手势识别

康来[1,2]，张亚坤[3]
(1. 国防科技大学 系统工程学院，湖南 长沙 410073；2. 国防科技大学 大数据与决策实验室，湖南 长沙 410073；
3. 中国人民解放军61081部队，北京 100089)

摘要：基于视觉的手势识别是虚拟现实、游戏仿真等领域常用的人机交互手段。在实际应用中，手势动作快速变化将导致传统RGB相机或深度相机成像模糊，给手势识别带来巨大挑战。针对上述问题，利用动态视觉传感器捕捉高速手势运动信息，*提出一种基于多维投影时空事件帧(spatiotemporal event frame, STEF)的动态视觉数据手势识别方法。将时空信息嵌入到数据投影面融合形成多维投影时空事件帧，克服现有动态视觉信息事件帧表达方法时域信息丢失的局限性，提升动态视觉传感数据的特征表达能力。*在此基础上，采用先进的脉冲神经网络对时空事件帧进行分类实现手势识别。在公开数据集上的识别精度达到96.67%，性能优于同类方法，表明该方法可显著提升动态视觉传感数据手势识别准确率。

关键词：动态视觉传感器；手势识别；多维投影；时空事件帧；脉冲神经网络
中图分类号：TP391 文献标志码：A 文章编号：1004-731X(2024)03-0649-10
DOI: 10.16182/j.issn1004731x.joss.23-0223
引用格式: 康来, 张亚坤. 基于多维投影时空事件帧的动态视觉传感手势识别[J]. 系统仿真学报, 2024, 36(3): 649-658.
Reference format: Kang Lai, Zhang Yakun. Gesture Recognition for Dynamic Vision Sensor Based on Multi-dimensional Projection Spatiotemporal Event Frame[J]. Journal of System Simulation, 2024, 36(3): 649-658.

Gesture Recognition for Dynamic Vision Sensor Based on Multi-dimensional Projection Spatiotemporal Event Frame

Kang Lai[1,2], *Zhang Yakun*[3]
(1. College of Systems Engineering, National University of Defense Technology, Changsha 410073, China; 2. Laboratory for Big Data and Decision, National University of Defense Technology, Changsha 410073, China; 3. PLA 61081 Troops, Beijing 100089, China)

Abstract: Vision-based gesture recognition is a commonly used means of human-computer interaction in the fields of virtual reality and game simulation. In practical applications, rapid changes in gesture movements will lead to blurred imaging with traditional RGB cameras or depth cameras, which brings great challenges to gesture recognition. To solve the above problems, *a dynamic visual data gesture recognition method based on a multi-dimensional projection spatiotemporal event frame (STEF) is proposed by a using dynamic vision sensor to capture high-speed gesture movement information. The spatiotemporal information is embedded in the data projection surface and fused to form a multi-dimensional projection STEF, which overcomes the limitation of the time-domain information loss of the existing event frame expression method of dynamic visual information and improves the feature expression ability of dynamic visual sensing data.* On this basis, advanced spiking neural networks are used to classify STEFs to realize gesture recognition. The recognition accuracy of the above method on the public dataset reaches 96.67%, which is better than similar methods, indicating that the proposed

收稿日期：2023-02-28 修回日期：2023-04-23
基金项目：国家自然科学基金(61873274)
第一作者：康来(1983-)，男，副教授，博士，研究方向为计算机视觉与模式识别、虚拟现实技术。E-mail：kanglai@nudt.edu.cn

method can significantly improve the accuracy of gesture recognition in dynamic visual sensing data.
Keywords: dynamic vision sensor; gesture recognition; multi-dimensional projection; spatiotemporal event frame; spiking neural network

0 引言

传统相机传感器(如：CMOS、CCD)以特定频率拍摄获取图像帧，每秒钟拍摄的图像数量称为帧率。普通消费级相机的帧率通常为30~60 帧/s，由于曝光需要一定时间，成像通常存在数十毫秒甚至更大的延时。如果曝光期间物体高速运动，则会产生成像模糊。此外，传统相机的动态范围较低(约 60 dB)，在光线极暗或者亮度极高的环境下，均无法获取足够的信息。上述问题限制了传统相机在快速运动或对实时性要求极高的场景中的应用。

事件相机(event camera)是一种生物启发(bio-inspired)的神经拟态(neuromorphic)新型视觉传感相机，能够响应局部亮度变化，也称作动态视觉传感器(dynamic vision sensor, DVS)[1]。与传统相机使用快门捕获图像不同，DVS 中的每个像素都独立、异步地运行，仅在亮度变化(变亮或变暗)达到一定阈值时才输出一个事件(event)，这些事件称为动态视觉传感数据。与传统相机相比，DVS 具有诸多优势。首先，DVS 的延迟极低(通常为微秒级)，可以更快地捕捉到亮度的变化，且不会像传统相机那样产生运动模糊。其次，由于动态视觉传感器异步监测每个像素，只有当强度变化超过阈值时才会产生事件，在没有变化的情况下始终处于静默状态，因此具有更低的能耗。此外，DVS 的动态范围远高于传统相机，通常可超过 140 dB。目前，DVS 技术在学术界逐渐引起了较大关注，相关研究已经拓展到目标跟踪[2]、目标识别[3]、手势识别[1]、结构光三维扫描[4]、光流估计[5]、高动态范围(high dynamic ranging, HDR)图像重建[6]、即时定位和建图(simultaneous localization and mapping, SLAM)[7-8]等领域。

由于DVS的工作模式及其数据形态均与传统相机完全不同，为了更好的发挥DVS的固有优势，需要设计全新的数据表达和数据处理方法[9]。在DVS数据表达方面，需针对特定应用将事件数据转换为其他表达形式，以便提取出有用信息和特征。常见的数据表达形式包括事件帧(event frame)、时间表面(time surface)、体素网格(voxel grid)、三维点集(3D point set)等形式[9]。其中，事件帧将一组事件数据转换为一幅图像，该表达形式在文献中占据重要地位[10-11]，其主要原因是它可通过简单的变换(比如：统计各像素位置的事件数量[12])将事件数据转换为标准、易于处理的二维图像数据。事件帧保留了丰富的场景边界信息，而且可以利用现有基于图像的计算机视觉算法对其进行处理。事件帧方法的不足是抛弃了事件数据的稀疏性[13]，且丢失了时域信息。时间表面也将事件数据转换为二维图像，但各像素记录的是相应位置上事件发生的最近时刻[14]。时间表面图本质上反映了各像素位置的历史变化情况，值越大说明该位置发生变化的时间越近。时间表面法通常结合归一化处理不同的变化速度[15]，采用滤波方法减少噪声的影响[16]。三维点集表达将事件数据视为几何数据，3 个维度分别对应于二维像素坐标和事件时刻[17]。DVS 数据处理方式很大程度上取决于数据表达方式。对于原始事件数据，通常采用卡尔曼滤波和粒子滤波等概率滤波器(贝叶斯方法)进行处理。滤波器能够自然的处理异步数据，具有较低的时延，而且能够融合多个传感器的数据[18]。另一类主流处理方法是基于人工神经网络(artificial neural network, ANN)的方法，包括卷积神经网络(convolutional neural network, CNN)[1,17]和脉冲神经网络(spiking neural network, SNN)[19-20]。DVS 数据表达和处理方法目前尚没有统一的标准

DOI: 10.16182/j.issn1004731x.joss.23-0223

和框架，DVS 技术及相关应用仍处于快速发展阶段，相关探索具有巨大的研究价值和空间，相关综述可参考文献[9]和[21]。

本文关注基于动态视觉传感数据的手势识别问题。手势识别是一种重要的人机交互手段，被广泛应用于虚拟现实、游戏仿真等领域[17]。由于普通消费级 RGB 相机或深度相机曝光通常需要数十毫秒时间，手势动作快速变化时，曝光期间场景动态变化导致成像模糊。本文采用动态视觉传感器可以获得微秒级极低延迟事件数据，从根本上避免成像模糊问题，并设计新的数据处理和识别方法实现手势识别。本文提出一种基于多维投影时空事件帧(spatiotemporal event frame, STEF)的动态视觉数据手势识别方法，将时间信息嵌入到数据投影面，与原单通道空间事件帧进行融合，增强动态视觉传感数据的特征表达能力，然后采用先进的脉冲神经网络结合自动精度混合模式实现动态视觉数据手势识别。最后，在公开数据集上对本文方法进行测试评估。

1 本文方法

事件相机与传统相机相比具有诸多优点，但由于动态视觉传感器事件流数据与传统 RGB 图像数据形态之间巨大差异，传统的图像处理算法均无法直接对动态视觉传感数据进行处理，因此需要对事件流数据进行某种形式的转换和表达。本文动态视觉传感数据手势识别基本流程如图 1 所示。

1.1 事件相机成像模型

传统相机拍摄获取的图像记录的是某一时刻场景的明暗颜色信息，图像所有像素均在相同的时刻记录。根据硬件性能不同，传统 RGB 相机图像分辨率一般在百万至千万像素级别，每秒可以拍摄的图像数量在几十帧到几千帧。因此，传统相机在空间上具有高分辨率，而在时域上分辨率较低。

事件相机记录场景位置亮度的变化情况，只有当亮度变化值超过一定阈值才会触发一个事件。随着场景变化，事件源源不断的产生输出事件流。每个像素是相互独立的，数据处理异步执行，事件刷新延时通常为微秒级。因此，事件相机的时域分辨率远高于传统 RGB 相机，可以捕捉到快速变化的细节信息。每个事件表示为

$$e=(x,y,t,p) \tag{1}$$

式中：x、y为二维图像平面内像素的横、纵坐标；t为事件发生的时刻；$p\in\{+1,-1\}$为事件的极性。记$e_k=(x,y,t_k,p_k)$为t_k时刻像素位置(x,y)发生的事件，$I(x,y,t)$为t时刻像素位置(x,y)的图像亮度，Δt为t时刻之后逝去时间，$C>0$为时序亮度阈值，一旦像素亮度满足：

$$p_{k+1}\cdot\left[\ln\left(I(x,y,t_k+\Delta t_k)\right)-\ln\left(I(x,y,t_k)\right)\right]\geqslant C \tag{2}$$

则将触发产生一个新的事件：

$$e_{k+1}=(x,y,t_k+\Delta t_k,p_{k+1}) \tag{3}$$

从上述模型原理可以看出，事件相机仅记录场景的变化，因此又称为动态视觉传感器。

事件相机成像基本原理示意图和相机数据示例如图 2 所示。

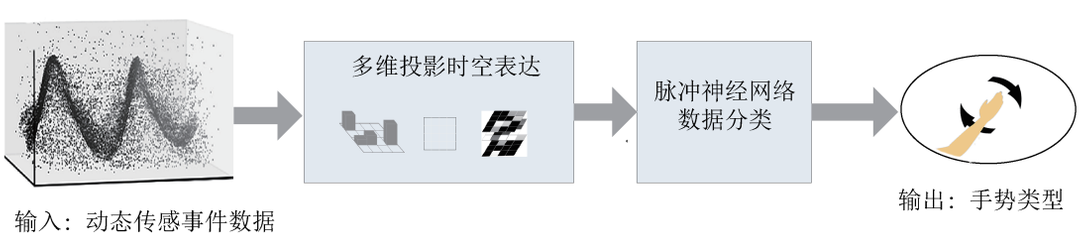
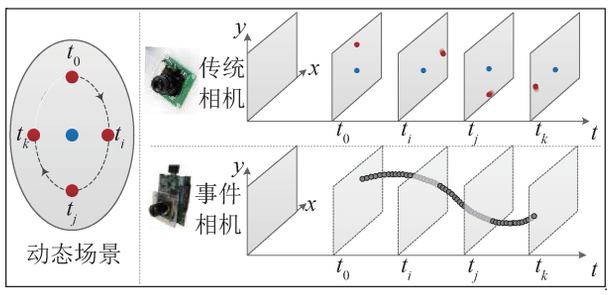
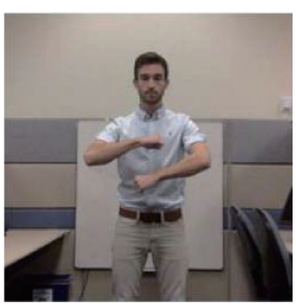


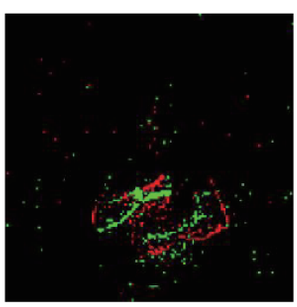
图 1 动态传感数据手势识别基本流程
Fig. 1 Workflow of gesture recognition based on dynamic sensing data



(a) 成像模型示意图



(b) RGB图像示例[1]



(c) 事件流示例[1]

图 2　相机成像模型及图像示例
Fig. 2　Illustration of camera model and images

图2(a)中展示的动态场景包含2个圆点，外侧圆点以内部圆点为中心顺时针高速旋转，动态场景中标记了t_0,t_i,t_j,t_k 4个时刻圆点所在位置。传统相机在t_0,t_i,t_j,t_k 4个时刻分别对上述动态场景进行拍摄获得的4幅图像如图2(a)右上子图所示，各图像中内部圆点清晰，而外侧圆点由于快门速度不够出现“拖尾”现象。利用动态视觉传感器拍摄上述场景产生的事件流如图2(a)右下子图所示，每个灰色圆点代表一个事件，包含了像素坐标、事件时刻(注意，图中未区分事件的极性)。从图2(a)可以看出，动态视觉传感器产生的事件流数据和传统RGB相机的二维图像数据形态有本质区别。为了进一步直观地理解动态视觉传感数据流形态，图2(b)和(c)分别展示了两类相机拍摄DVS128 Gesture数据集中“arm roll”手势场景获取的数据示例[1]。其中，图2(c)包含了某个5 ms时间范围内的所有事件，采用红(正极，即p=1)、绿(负极，即p=−1)两种颜色区分事件极性，从图中可以看出运动区域的边界信息，但也包含明显的背景噪声。

1.2 多维投影事件数据表达

本文提出基于多维投影的动态传感数据时空事件帧STEF表达，可以有效克服传统积分帧抽取导致的时域信息丢失问题，提高数据的表达能力，进而达到提升手势识别性能的目的。

1.2.1 归一化积分事件帧抽取方法

积分帧的基本思想是将某段时空邻域中的所有事件进行累加获得各像素位置的事件数量(区分不同极性)，并将结果转化为二维图像。其优点是转化为数据积分帧后可以直接采用现有图像处理算法对其进行处理，既保持了动态视觉传感器的高动态性，又极大的简化了数据处理流程。

事件流数据积分帧抽取过程示意图如图3所示。记积分帧为$\mathcal{F}_t \in \mathbb{R}^{W\times H\times 2}$，它包含两幅尺寸为$W\times H$像素的图像，分别对应于正、负极性事件积分帧。积分帧特定像素位置(x,y)的像素值记为$\mathcal{F}_t(x,y,p)$。其中，p与式(1)中的含义相同，均表示事件的极性，这里也用于区分正、负极事件积分帧。则事件集合为

$$\mathcal{E}=\left\{e_i=\left(x_i,y_i,t_i,p_i\right)|i=0,1,\cdots,N\right\} \tag{4}$$

式中：N为事件的数量。

对应的事件流数据积分帧$\mathcal{F}_t$各像素数值的计算公式为

$$\mathcal{F}_t(x,y,p)=\sum_{i=0}^{N}E(x,y,p,x_i,y_i,p_i) \tag{5}$$

其中，

$$E(x,y,p,x_i,y_i,p_i)=\begin{cases}1, (x,y,p)=(x_i,y_i,p_i)\\ 0, \quad \text{其他}\end{cases} \tag{6}$$

图3中，由于数据积分帧$\mathcal{F}_t$各像素的值取决于该像素位置事件的数量和极性，因此各积分帧$\mathcal{F}_t$的像素取值范围存在差异，甚至可能超出RGB通道的最大值255。为了对事件流积分数据帧进行规范化统一描述，将积分帧$\mathcal{F}$进行归一化处理，事件集合转换为两幅尺寸为$W\times H$像素的灰度图事件帧$\mathcal{H}_t \in \mathbb{R}^{W\times H\times 2}$，其像素值的取值范围为$\mathcal{H}_t(x,y,p)\in[0,1]$。

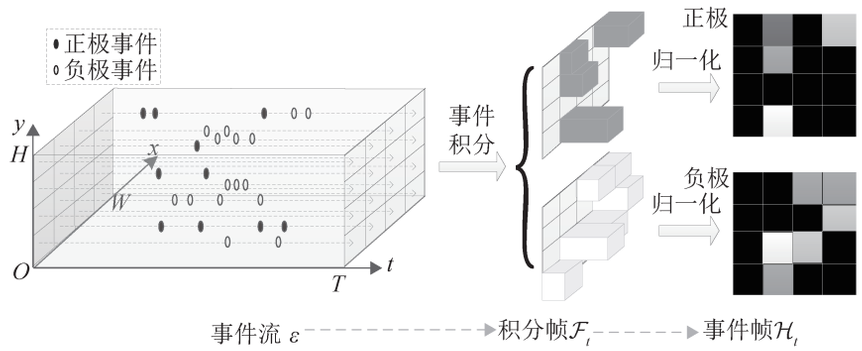


图 3 事件流数据积分帧提取过程示意图
Fig. 3 Integration frame extraction process of event stream data

1.2.2 多维投影时空事件帧抽取方法

如图3所示，归一化积分事件帧$\mathcal{F}_t$的提取过程本质上是沿时间轴t方向投影。由于积分帧抽取过程中事件的顺序无法保持，因此归一化积分事件帧不可避免的丢失了时域信息，削弱了动态视觉传感器对场景的描述能力。为了弥补上述缺陷，本文提出基于多维投影的时空事件帧STEF数据表达方式。

采用类似于获取$\mathcal{F}_t$的计算过程，沿动态视觉传感器坐标系x轴方向投影获取积分帧$\mathcal{F}_x \in \mathbb{R}^{T\times H\times 2}$。$\mathcal{F}_x$包含两幅尺寸为$T\times H$像素的图像，分别对应于$t-y$平面正、负极性事件积分帧。$\mathcal{F}_x$包含了事件的$y$坐标信息和时间信息$t$，因此对仅包含$x,y$坐标信息的积分帧$\mathcal{F}_t$是一种有效的增强和补充。更进一步，可沿$y$轴方向投影获取积分帧$\mathcal{F}_y \in \mathbb{R}^{T\times W\times 2}$。将$\mathcal{F}_t$、$\mathcal{F}_x$和$\mathcal{F}_y$进行对齐、融合以及归一化处理，可获得包含6幅灰度图(尺寸为$W\times H$像素)的时空事件帧$\mathcal{H} \in \mathbb{R}^{W\times H\times 6}$。这里，对齐操作主要是对$\mathcal{F}_x$和$\mathcal{F}_y$进行重采样，使其各通道图像尺寸分别由$T\times H$、$T\times W$统一转换为$W\times H$。时空事件帧$\mathcal{H}$各通道图像承载了事件数据在时间和纵、横图像空间3个维度的信息，能够有效提升事件帧表达能力。

1.2.3 时空事件帧集合构建方法

将事件集合$\mathcal{E}=\{e_i|i=0,1,\cdots,N\}$按照时间戳均匀划分为$Q$个事件子集$\mathcal{E}_s(s=0,1,\cdots,Q-1)$。事件集合$\mathcal{E}$中事件的起始时间戳为$t_0$，结束事件戳为$t_N$，各事件子集$\mathcal{E}_s$的划分起始时刻和结束时刻为

$$\begin{cases} t_{\text{start}}(s)=t_0+\left(\dfrac{t_N-t_0}{Q}\right)s \\ t_{\text{end}}(s)=t_0+\left(\dfrac{t_N-t_0}{Q}\right)(s+1) \end{cases} \tag{7}$$

记$N_{\text{start}}(s)$和$N_{\text{end}}(s)$分别为满足式(8)所示不等式条件的最小整数：

$$\begin{cases} t_{N_{\text{start}}(s)} \geqslant t_{\text{start}}(s) \\ t_{N_{\text{end}}(s)} \leqslant t_{\text{end}}(s) \end{cases} \tag{8}$$

则事件子集$\mathcal{E}_s$可以显式地表示为

$$\{e_k=(x_k,y_k,t_k,p_k)|k=N_{\text{start}}(s),N_{\text{start}}(s)+1,\cdots,N_{\text{end}}(s)\} \tag{9}$$

在此基础上，按照第1.2.2所述方法对$\mathcal{E}_s$进行多维时空投影处理，最终将事件集合$\mathcal{E}$转换为Q个时空事件帧集合：

$\{\mathcal{H}_s|\mathcal{H}_s \in \mathbb{R}^{W\times H\times 6};s=0,1,\cdots,Q-1\}$。

1.3 基于脉冲神经网络的手势识别方法

多维投影时空事件帧STEF与传统多通道图像表达形式一致，可以采用经典深度神经网络CNN进行训练和分类[1,17]，也可采用功耗更低、类生物结构合理性的新型脉冲神经网络SNN对上述数据进行处理。本文采用文献[20]提出的脉冲神经网络架构对时空事件帧集$\{\mathcal{H}_s\}$进行分类实现手势识别，该网络的基本架构如图4所示。在训练过程

中，同时训练SNN模型突触权重和膜电位时间参数，降低初值的敏感性及加速训练过程。

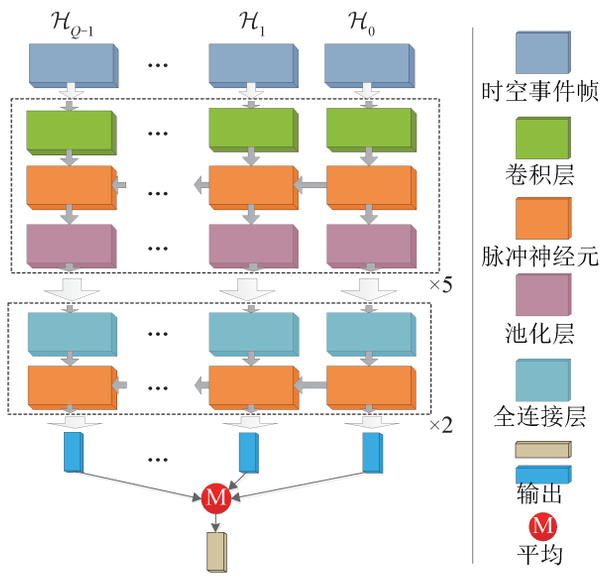


图 4　基于脉冲神经网络的手势识别网络架构
Fig. 4　Architecture of gesture recognition network based on spiking neural network

2　实验验证

为了验证本文方法的有效性，实现了相关算法并在公开数据集上对其性能进行评估。其中，深度学习训练和推理硬件为NVIDIA RTX 3090显卡，显存大小为24 GB，通用并行计算架构为NVIDIA CUDA平台，深度学习模型训练及推理基于SpikingJelly实现。其中，SpikingJelly是一个基于PyTorch的开源脉冲神经网络深度学习框架[20]。

2.1　数据集

本文使用的数据集为DVS128 Gesture公开数据集[1]。该数据集由iniLabs DVS128相机拍摄，图像空间分辨率为128×128像素，事件时间戳粒度为微秒刻度(即时域刷新频率为每秒100万帧)。该数据集包括29名对象在不同照明条件下(自然光、荧光灯和LED灯)的动作数据片段，23名对象数据为训练集、其余6名对象数据片段为测试集。手部和手臂动作姿势类型共11类(划分表如表1所示)。其中，第11类是除去前10类动作以外的其他动作类型，由于该类型手势数据不确定性更大，容易造成训练以及测试的不稳定性，对手势识别提出了更大的挑战。

表 1　DVS128 Gesture手势类型划分
Table 1　Gesture categories in DVS128 gesture

标签	动作类型
1	hand clap
2	right hand wave
3	left hand wave
4	right arm clockwise
5	right arm counter clockwise
6	left arm clockwise
7	left arm counter clockwise
8	arm roll
9	air drums
10	air guitar
11	other gestures

DVS128 Gesture数据集数据样例如图5所示。为了便于展示事件数据，图5中各图展示了5 ms时间范围内的所有事件，并且采用红、绿两种颜色区分事件极性。图5第1行为“hand clap”动作5个顺序时间段的事件数据，由于硬件性能和环境条件影响背景中产生一定量的噪声，人眼可大致分辨该动作类型。图5第2、3行分别展示了其他10种手势动作的事件数据。

2.2　实验设置

本文所有实验均采用统一的超参数设置，具体如下。训练轮数(epochs)为256，批大小(batchsize)为16。初始学习率(learning rate)为0.001，并采用余弦退火策略动态调整学习率[22]。在神经网络推理过程中，通过开启自动混合精度(automatic mixed precision, AMP)，对不同的网络层采用不同的数据精度进行计算，实现节省显存和加快训练速度的目的。此外，通过CuPy实现Numpy数组的并行加速运算。

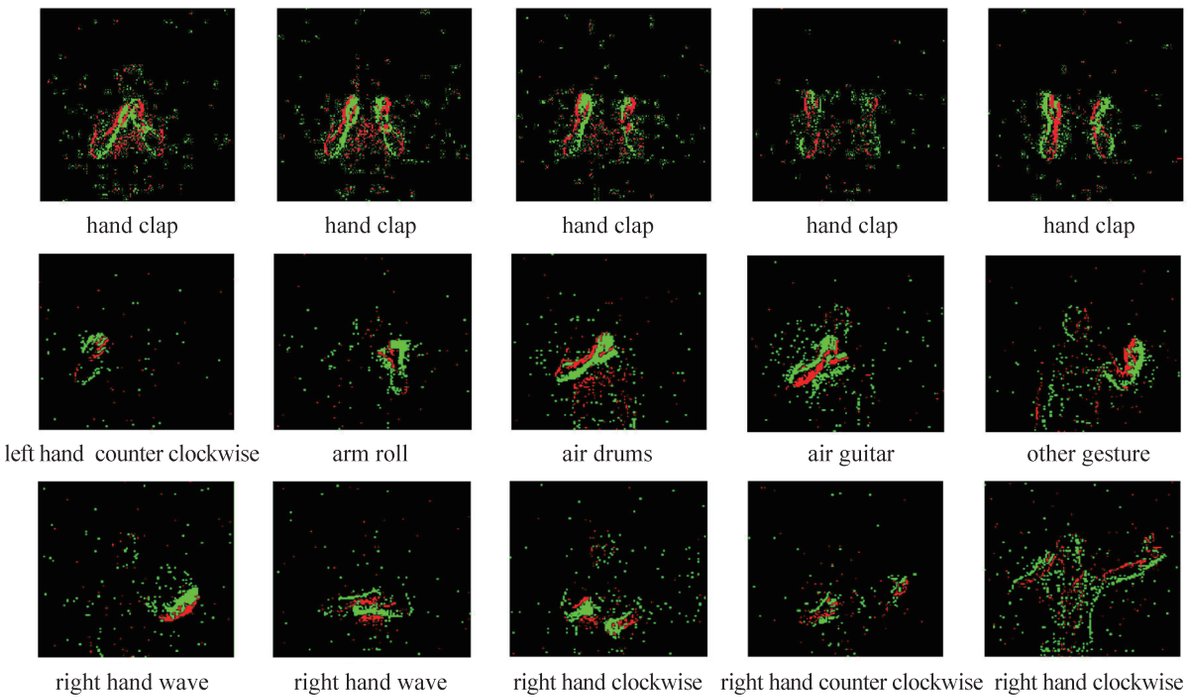
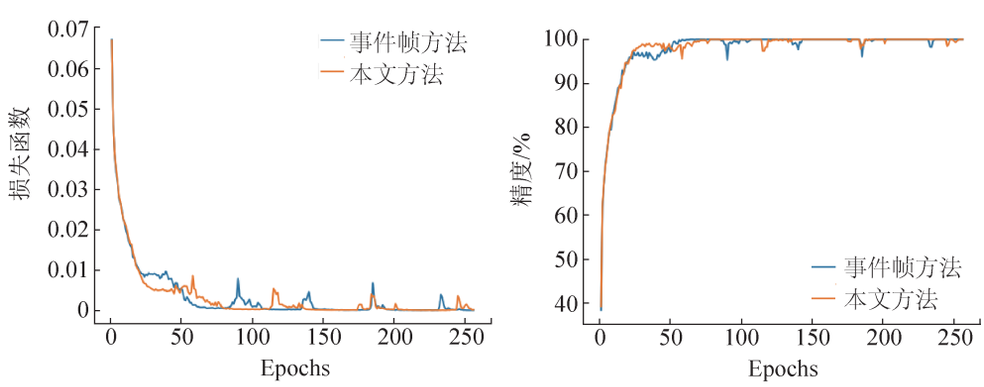


图 5 DVS128 Gesture 数据集示例
Fig. 5 Samples from DVS128 Gesture dataset

2.3 结果与分析

本文脉冲神经网络训练过程如图 6 所示。事件帧方法指的是采用归一化积分事件帧对事件数据进行处理(见第 1.2.1 节)，然后采用脉冲神经网络进行手势识别。从图 6 展示的训练过程可视化中可以看出，2 种方法具有相似的训练特性，本文方法的训练收敛速度略快，说明了本文方法引入的时空事件帧的有效性。在运算效率方面，本文实验使用 NVIDIA 3090 显卡进行训练，在未设置 Cupy 与自动混合精度运算时每轮训练耗时约 39 s，设置 Cupy 与自动混合精度，并对数据结构进行优化后，最终每轮训练耗时约 12 s，大幅降低了时间开销。

利用本文第 1.2.3 节所述方法为 DVS128 Gesture 数据集每个事件数据片段创建一个包含 16 个时空事件帧的时空事件帧集合。时空事件帧创建示例如图 7 所示(注意：本图仅展示正极性事件)。



(a) 损失函数值变化 (b) 手势识别精度变化

图 6 脉冲神经网络训练过程
Fig. 6 Training process of spiking neural network

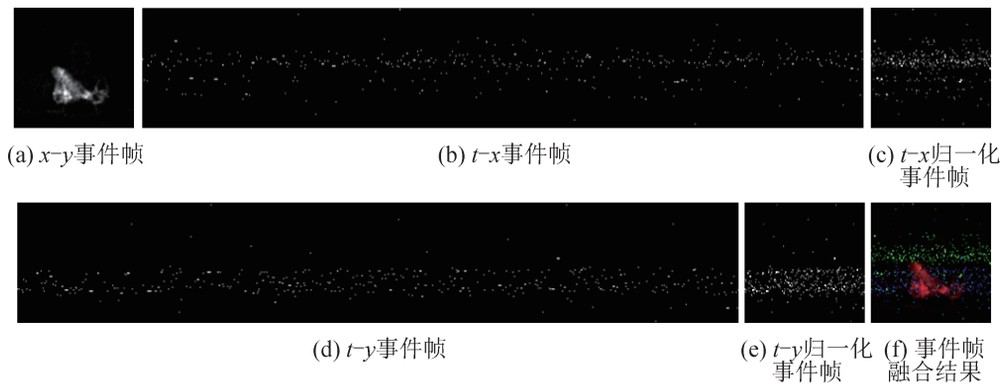


图 7　本文时空事件帧创建示例
Fig. 7　Illustration of construction of proposed STEF

其中，图 7(a)为沿时间轴 t 投影获得的 $x-y$ 平面内的事件帧，图 7(b)为沿 y 轴投影获得的 $t-x$ 平面内的事件帧，图 7(c)为 $t-x$ 平面事件帧归一化处理结果，图 7(d)为沿 x 轴投影获得的 $t-y$ 平面内的事件帧、图 7(e)为 $t-y$ 平面事件帧归一化处理结果。将各投影维度事件帧融合后可获得包含 6 个通道的时空事件帧。图 7(f)展示了图 7(a)、(c)、(e)的融合结果。为便于观察，图 7(f)仅展示正极性事件，并将 t、y、x 轴对应的事件帧作为颜色通道进行映射，实际计算过程种无需生成 RGB 图像。

图 8 为不同方法在 DVS128 Gesture 测试集上的手势识别精度对比结果。其中，Low power 方法将动态视觉数据转换为传统空间事件帧，然后采用传统 CNN 模型进行识别，其识别精度为 94.59%。Event cloud 方法将事件数据作为点云进行处理和识别，其识别精度为 95.32%。其余方法均基于传统空间事件帧数据表达，采用 SNN 进行识别，精度在 92.10%~96.18% 之间。SpikingJelly (AMP)为文献[20]所述方法在开启自动混合精度且每个手势片段划分为 16 的实验结果，在相同的参数设置下，本文方法的识别精度达到了 96.67%。上述实验结果表明，本文方法可以显著提升动态视觉传感数据手势识别的准确率，性能优于同类方法。与 SNN 传统方法实验相比，本文方法的优势主要体现在对随机手势(表 1 中的动作类型 11)这类较困难的手势具有更好的识别效果，而误识别主要集中在多种光源条件下更具挑战的复杂情形。

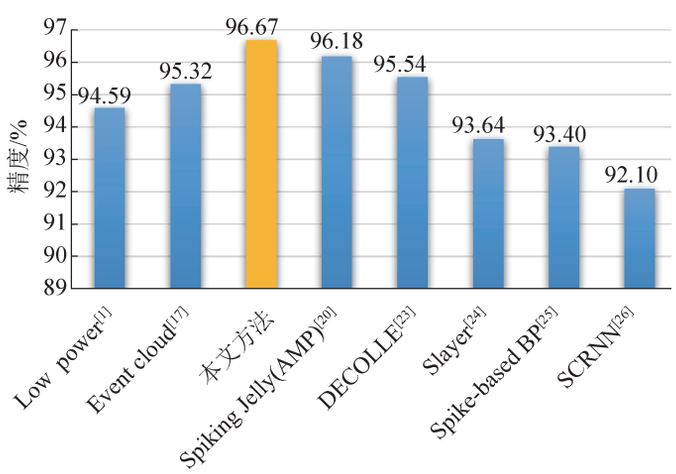


图 8　手势识别精度比较
Fig. 8　Comparison of gesture recognition accuracy

DOI: 10.16182/j.issn1004731x.joss.23-0223

3 结论

本文提出一种基于多维投影时空事件帧的动态视觉数据手势识别方法，将时间信息嵌入到数据投影面，与原单通道空间事件帧进行融合，有利于增强动态视觉传感数据的特征表达能力，采用先进的脉冲神经网络结合自动精度混合模式实现动态视觉数据手势识别。在公开数据集上的测试表明，本文方法的性能优于同类算法。后续研究拟尝试多种神经网络模型以验证本文STEF方法的适应性与普遍性，并针对多种光源条件引起的误识别情况，采用滤波方式降低背景噪声对识别性能的影响。从而进一步发挥动态视觉传感器高动态、低冗余、低消耗以及不受光照条件影响的优势，为虚拟仿真、智能驾驶、互动游戏等应用提供人机交互支持。

参考文献：

[1] Amir A, Taba B, Berg D, et al. A Low Power, Fully Event-based Gesture Recognition System[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2017: 7388-7397.

[2] Arren Glover, Chiara Bartolozzi. Event-driven Ball Detection and Gaze Fixation in Clutter[C]//2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, NJ, USA: IEEE, 2016: 2203-2208.

[3] Orchard G, Cedric Meyer, Etienne-Cummings R, et al. HFirst: A Temporal Approach to Object Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(10): 2028-2040.

[4] Matsuda N, Cossairt O, Gupta M. MC3D: Motion Contrast 3D Scanning[C]//2015 IEEE International Conference on Computational Photography (ICCP). Piscataway, NJ, USA: IEEE, 2015: 1-10.

[5] Zhu Zihao, Yuan Liangzhe, Chaney K, et al. EV-FlowNet: Self-supervised Optical Flow Estimation for Event-based Cameras[EB/OL]. (2018-08-13) [2023-01-07]. https://arxiv.org/abs/1802.06898.

[6] Henri Rebecq, René Ranftl, Koltun V, et al. High Speed and High Dynamic Range Video with an Event Camera [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(6): 1964-1980.

[7] Henri Rebecq, Timo Horstschaefer, Guillermo Gallego, et al. EVO: A Geometric Approach to Event-based 6-DOF Parallel Tracking and Mapping in Real Time[J]. IEEE Robotics and Automation Letters, 2017, 2(2): 593-600.

[8] Antoni Rosinol Vidal, Henri Rebecq, Timo Horstschaefer, et al. Ultimate SLAM? Combining Events, Images, and IMU for Robust Visual SLAM in HDR and High-speed Scenarios[J]. IEEE Robotics and Automation Letters, 2018, 3(2): 994-1001.

[9] Guillermo Gallego, Tobi Delbrück, Orchard G, et al. Event-based Vision: A Survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(1): 154-180.

[10] Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, et al. Event-based Vision Meets Deep Learning on Steering Prediction for Self-driving Cars[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 5419-5427.

[11] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, et al. EKLT: Asynchronous Photometric Feature Tracking Using Events and Frames[J]. International Journal of Computer Vision, 2020, 128(3): 601-618.

[12] Liu Min, Delbruck T. Adaptive Time-slice Block-Matching Optical Flow Algorithm for Dynamic Vision Sensors[C]//British Machine Vision Conference (BMVC) 2018. UK: BMVC, 2018: 168589.

[13] Alessandro Aimar, Hesham Mostafa, Enrico Calabrese, et al. NullHop: A Flexible Convolutional Neural Network Accelerator Based on Sparse Representations of Feature Maps[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(3): 644-656.

[14] Xavier Lagorce, Orchard G, Francesco Galluppi, et al. HOTS: A Hierarchy of Event-based Time-surfaces for Pattern Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(7): 1346-1359.

[15] Jacques Manderscheid, Amos Sironi, Nicolas Bourdis, et al. Speed Invariant Time Surface for Learning to Detect Corner Points with Event-based Cameras[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2019: 10237-10246.

[16] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, et al. HATS: Histograms of Averaged Time Surfaces for Robust Event-based Object Classification[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 1731-1740.

[17] Wang Qinyi, Zhang Yexin, Yuan Junsong, et al. Space-time Event Clouds for Gesture Recognition: from RGB Cameras to Event Cameras[C]//2019 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway, NJ, USA: IEEE, 2019: 1826-1835.

[18] Guillermo Gallego, Jon E A Lund, Elias Mueggler, et al. Event-based, 6-DOF Camera Tracking from Photometric Depth Maps[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(10): 2402-2412.

[19] Roy K, Jaiswal A, Panda P. Towards Spike-based Machine Intelligence with Neuromorphic Computing[J]. Nature, 2019, 575(7784): 607-617.

[20] Fang Wei, Yu Zhaofei, Chen Yanqi, et al. Incorporating Learnable Membrane Time Constant to Enhance Learning of Spiking Neural Networks[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2021: 2641-2651.

[21] 黄铁军, 余肇飞, 李源, 等. 脉冲视觉研究进展[J]. 中国图象图形学报, 2022, 27(6): 1823-1839.
Huang Tiejun, Yu Zhaofei, Li Yuan, et al. Advances in Spike Vision[J]. Journal of Image and Graphics, 2022, 27 (6): 1823-1839.

[22] Ilya Loshchilo, Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts[C]//ICLR 2017 (5th International Conference on Learning Representations). New York, USA: ICLR, 2017: 1-16.

DOI: 10.16182/j.issn1004731x.joss.23-0223