论文阅读：Events-Only 3D Human Pose Tracking with Spiking Spatiotemporal Transformer

前言

本文是某种三维扫描及重建技术，细枝末节较多，源代码地址如下。

```
git clone https://codeberg.org/ybh1998/EventPS.git
```

总览

首先，看完标题、摘要和结论，可了解到如下信息：

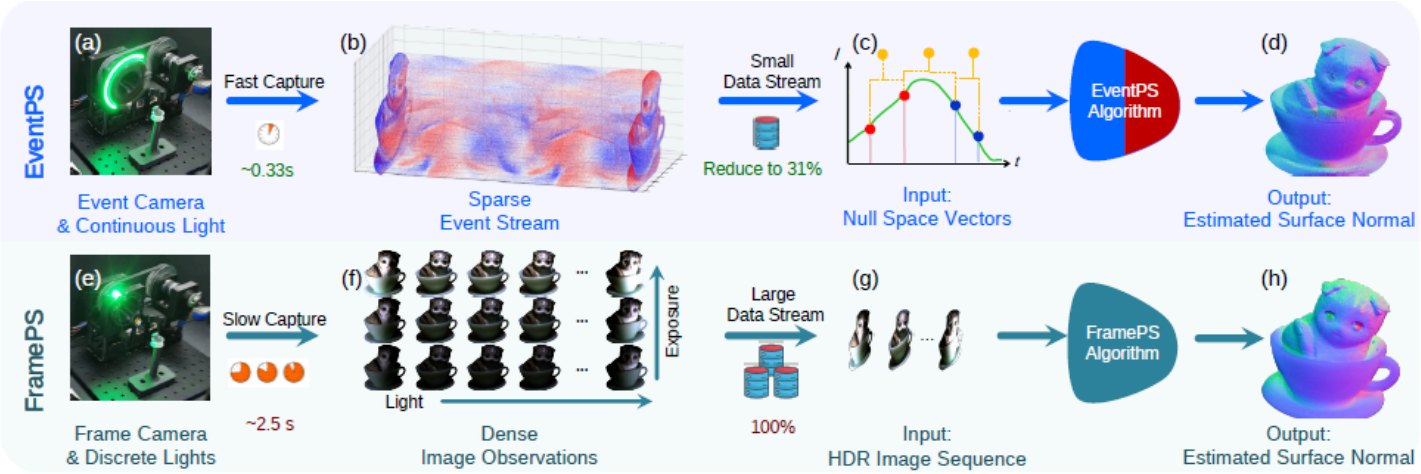
- i. 传统的立体光学法([Photometric Stereo](三维扫描仪 - 维基百科，自由的百科全书 (wikipedia.org)))为获取物体表面法向量图并由此计算深度图，需要在不同光照条件下捕获多个高动态范围图像，难以实时应用。
- ii. 事件相机(Event Camera)仅根据图片像素辐射变化来估计法线，能有效提高数据利用效率。本文因此提出了一种基于深度相机的实时立体光学法，融合了基于优化和基于深度学习的立体光学技术，且能适应非朗伯表面情形[Lambertian](朗伯余弦定律 - 维基百科，自由的百科全书 (wikipedia.org))并提高了噪声鲁棒性，在动态场景的实时传感和法线测量上显示出巨大潜力。

一、引言

由于实际场景中非朗伯表面带来的附加阴影(物体表面由于自身几何形状的遮挡而形成的阴影区域。这种阴影通常出现在物体的凹陷部分或背光侧，是物体表面的一部分被其他部分遮挡而无法直接接收到光源照射所致)、镜面反射和各种噪声等问题，使用传统的立体光学法来实现鲁棒性较强的物体表面法向量估计较为繁琐耗时。而事件相机尽管有高时间分辨率(在给定时间段内能够捕捉和标记事件的频率和精度)、高亮度变化范围、低带宽等优势，但由于其所摄像素本质上是辐射值变化量而不是辐射绝对值，所以需要立体视觉技术针对其重新定制。为此，作者做出了如下开创性贡献：

- i. 提出了一种针对事件相机的根据图像辐射连续变化来估计物体表面法线的光度立体法EventPS
- ii. 提出将EventPS与基于优化和基于深度学习的方法集成来处理朗伯表面和非朗伯表面。
- iii. 建立了一个带有高速旋转光源的验证平台进行成果展示。

EventPS和传统的立体光学法具体区别如下图所示。



二、相关工作

1.基于深度学习的立体光学技术

主要分为全像素和每像素两种方法。全像素代表方法PS-FCN使用全卷积网络来处理输入图像，网络结构能够保留空间信息并从输入图像中直接预测每个像素的表面法线，避免了传统方法中复杂的几何计算和优化过程。每像素方法注重提取每个像素的局部特征，通过观察不同光照方向下的像素来估计表面法线。 目前两种方法及优化方案仍然需要不同照明条件下的大量图像，需要优化物体观察过程。高速相机(成本高)、多光谱相机(鲁棒性差)等成像系统不能解决问题。

2.基于事件相机的三维重建技术

事件摄像机检测场景中的辐射变化，这可能是由摄像机/物体移动或照明变化引起的。相关研究可分为两类：基于运动的方法和基于主动照明的方法。

- i. 基于运动的方法：将单个事件视为射线，以估计半密集 3D 结构(只对图像中的显著区域或具有丰富纹理的区域进行三维重建)和来自具有已知轨迹的事件相机的对象网格。
- ii. 基于主动照明的方法：通过投射已知图案（如条纹、点阵、网格等）到物体表面，然后通过摄像机捕捉这些图案的变形并最大化投影和事件相机之间的时空相关性来进行深度探测。

三、训练及实验结果

1.问题重述

事件相机模型。事件摄像机以对数尺度捕捉场景辐射变化。每个像素异步测量亮度变化。当像素x处的对数辐射亮度变化达到触发阈值C时，将触发事件{x, p, t}，其中t为时间戳，p ∈ {−1, +1}为极性，表示辐射率的减少或增加。假设短时间内在像素x处总共触发了K个事件。这些事件表示为 Ex ={x, pk, tk}，其中 k = {1, 2, ..., K}。像素 x 中的亮度值从 tk−1 到 tk 的变化为

$$\log(I_{\mathbf{x}}(t_k) + \epsilon) = \log(I_{\mathbf{x}}(t_{k-1} + \eta) + \epsilon) + p_k C$$

其中ε是一个小偏移值，以避免取对数为零，η是像素的不应时间[6]，均可省略。得到下式。

$$I_{\mathbf{x}}(t_k) = \exp(p_k C) \cdot I_{\mathbf{x}}(t_{k-1}).$$ 而假设一个物体被理想的远距离光源照亮，光源的辐射亮度是恒定的，方向被描述为归一化照明矢量函数 L(t)。对于图像坐标 x = (x, y) 处的像素，法向量为 nx，漫反射反照率 ax，在朗伯假设下，该像素的反射辐射亮度 ˆ Ix(t) 为：

$$\hat{I}_{\mathbf{x}}(t) = \max\left[0, a_{\mathbf{x}}(\mathbf{n}_{\mathbf{x}} \cdot \mathbf{L}(t))\right].$$ 代入前式得到

$$\max\left[0, a_{\mathbf{x}}(\mathbf{n}_{\mathbf{x}} \cdot \mathbf{L}(t_k))\right] = \exp(p_k C) \cdot \max\left[0, a_{\mathbf{x}}(\mathbf{n}_{\mathbf{x}} \cdot \mathbf{L}(t_{k-1}))\right].$$ 所以给定像素 x 处捕获的事件 Ex 和照明方向 L(t)，我们的目标是找到函数 f 来估计像素 x 处的表面法线使 ˆ nx 使其尽可能接近真实法线nx。

2.EventPS模型

对于事件相机捕获的静态朗伯表面对象，作者观察到之前在事件相机模型中设置的触发产生的事件信号有三个特性：

- 1.反照率不变性。事件信号与表面反照率轴ax无关。这意味着在光照方向发生相同变化的情况下，表面反照率ax不会影响事件触发。所以可将之前的模型进一步简化为

$$\max\left[0, \mathbf{n}_{\mathbf{x}} \cdot \mathbf{L}(t_k)\right] = \exp(p_k C) \cdot \max\left[0, \mathbf{n}_{\mathbf{x}} \cdot \mathbf{L}(t_{k-1})\right].$$

- 2.附加阴影中没有事件。Ix(tk) 在 tk 处的导数必定不为零。否则，不会触发任何事件。这一属性说明事件信号不包含附加阴影区域中像素的冗余信息，并且在任何事件时间戳 tk 处 I 应大于 0。因此，我们从等式两边删除 max 运算符。 $$\mathbf{n}_{\mathbf{x}} \cdot \mathbf{L}(t_k) = \exp(p_k C)(\mathbf{n}_{\mathbf{x}} \cdot \mathbf{L}(t_{k-1}))$$

- 3.线性无关零空间向量。对于每个像素，我们将每对连续事件信号转换为位于该像素处的物体表面切平面上的向量，该向量垂直于表面法线。我们将这些向量称为零空间向量，表示为 zk，其中 k = {1, 2, ..., K − 1}。 $$\mathbf{z}_k = \mathbf{L}(t_{k+1}) - \exp(p_{k+1} C)\mathbf{L}(t_k).$$ 显然在每个像素点上，该向量垂直于表面法线。 表面法线具有唯一性。如果所有的零空间向量都是线性相关的（即它们在同一条直线上），那么将会有无限多个表面法线向量与这些零空间向量垂直。这是因为在同一条直线上的向量没有提供足够的独立信息来唯一确定一个法线向量。 通过扫描模式引入曲线并计算法线。使用凸曲线作为扫描模式可以确保在每次扫描中获得线性无关的零空间向量。以下是其工作原理：

- i. **凸曲线的性质**：凸曲线上的任意三点不共线。这意味着通过扫描凸曲线，可以获得多个线性无关的点。
- ii. **线性无关的零空间向量**：从这些线性无关的点中，可以计算出多个线性无关的零空间向量。这些向量提供了足够的信息来唯一确定表面法线。

3.基于优化的EventPS

对于每个像素，作者将所有零空间向量组合成一个 3 × (K − 1) 矩阵 Zx。理论上，至少需要 3 个事件才能获得用于表面法线估计的 2 阶矩阵 Zx。给定足够的事件（即 K > 3），作者定义估计表面法线 ^ nx 的优化目标，以最小化以下均方误差 (MSE)。

$$\underset{\hat{\mathbf{n}}_\mathbf{x}}{\operatorname{argmin}} \|\mathbf{Z}_\mathbf{x}^\top \hat{\mathbf{n}}_\mathbf{x}\|_2.$$

这个优化问题是通过SVD来解决的，具体原理如下图。我们计算矩阵 ZxZx⊤ 的最小特征值对应的特征向量，然后得到表面法线 ˆ nx。我们将此方法命名为事件光度立体优化(EventPS-OP)。

最小二乘问题是优化问题中的一个基本问题，目标是找到一个向量 x，使得 $Ax \approx b$ 的误差最小。具体来说，目标是最小化以下目标函数：

$$\min_x \|Ax - b\|_2^2$$

通过SVD分解 $A = U\Sigma V^T$，我们可以将问题转化为更简单的形式：

$$\min_x \|U\Sigma V^T x - b\|_2^2$$

由于 U 是正交矩阵，可以通过左乘 U^T 来简化问题：

$$\min_x \|\Sigma V^T x - U^T b\|_2^2$$

设 $y = V^T x$ 和 $c = U^T b$，问题变为：

$$\min_y \|\Sigma y - c\|_2^2$$

由于 Σ 是对角矩阵，求解这个问题变得更加简单。最终，通过求解 y 并回代 $x = Vy$，可以得到最优解。

添加阈值来过滤掉最亮区域（最有可能在镜面高光中）和最暗区域（最有可能在附加/投射阴影中）可以有效提高立体光学技术的精度。在EventPS中，由于缺乏绝对辐射度信息，我们很难在事件信号中添加这样的阈值。然而，当观察到高对比度的强度变化时，事件会以高频率触发。在 光度立体设置中，当点穿过阴影边界（包括攻击阴影和投射阴影）或镜面高光时，通常会发生这种情况。通过对事件触发频率设置阈值，我们可以达到与频域最小二乘法添加阈值类似的目标。滤波后的零空间向量 ˆ Z 为

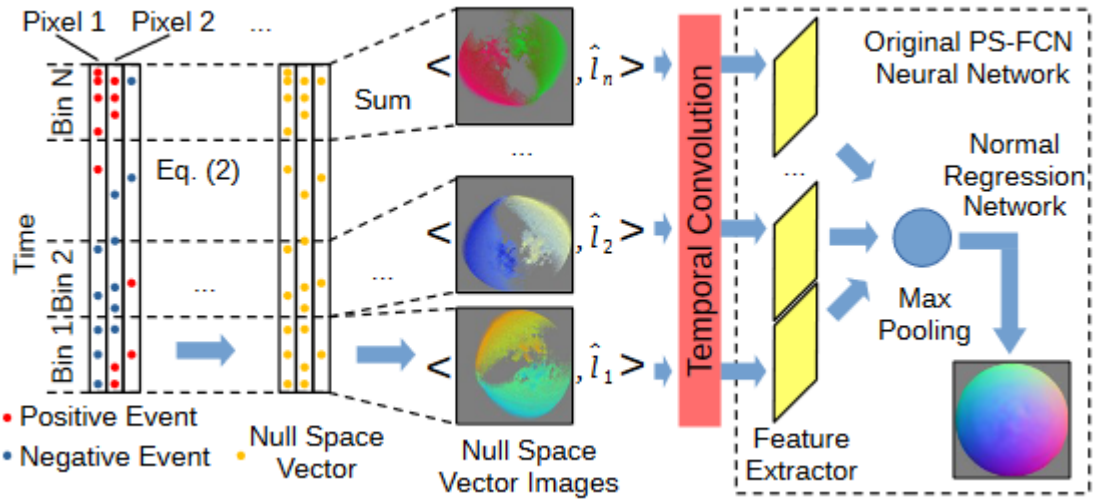
$$\hat{\mathbf{Z}} = \{\mathbf{z}_k \mid k > 1 \text{ and } t_k > t_{k-1} + \delta\}$$

4.基于深度学习的EventPS

作者采用了两种深度学习框架PS-FCN(全像素)和CNN-PS(每像素)并将其调整以适应事件信号的模态。

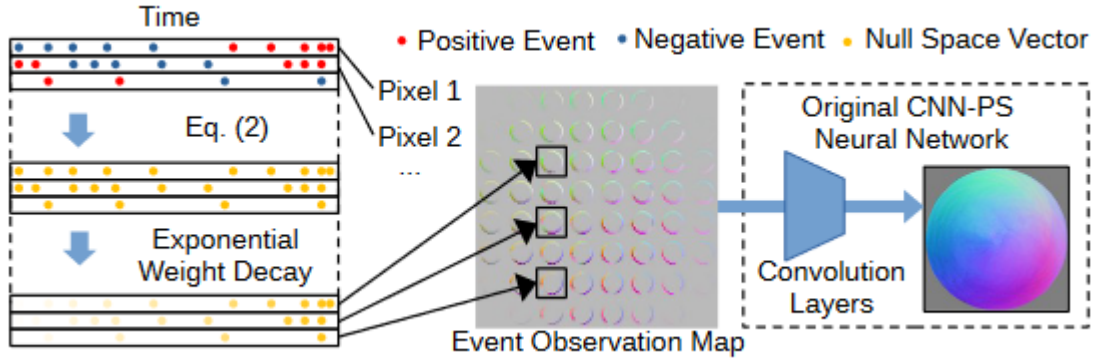
EventPS-FCN

原始的 PS-FCN 将卷积层应用于特定光照下的每个单独图像，并通过最大池化合并多个图像特征。作者通过构建零空间矢量图像作为输入来维持像素内关系，从而使 PS-FCN 适应事件模态（称为 EventPS-FCN）。首先将感兴趣的扫描时间段（通常是整个圆）划分为 N 个区间。使用方程式将事件转换为零空间向量。零空间矢量图像是通过将在每个时间单位每个像素中所有共享光线方向变化的零空间矢量相加而形成的。 通过向每个零空间矢量图像添加光方向 ˆ Li 通道来进行特征提取。由于事件特征比图像特征稀疏得多，并且相邻时间隙之间的差异不明显，因此作者添加两个时间卷积层来从相邻时间箱的事件中提取时间特征。然后将所有时间隙的特征最大池化在一起以估计表面法线。



EventPS-CNN

原始的 CNN-PS 通过从每个像素提取 32×32 观察图并在这样的观察图上应用卷积层来单独处理每个像素。类似地，使用所提出的 EventPS 公式从事件信号到零空间向量的转换也是在每个像素的基础上执行的。在事件观察图中，作者将通道数从 1（灰度图像）增加到 3（零空间向量的 x、y、z 轴）。每个像素代表相应照明方向的零空间矢量。通过这种方式，每个像素处的所有零空间向量都聚集在此事件观测图中，并馈送到原始 CNN-PS 模型。这种方法保留的零空间向量信息更多。



实验

与之前的立体光学法比较

在DiLiGenTEv数据集上测试。随着输入图像数量的增加，FramePS 显示数据速率呈线性增加，同时伴随着正常 MAE 的减少。相比之下，所提出的 EventPS 具有恒定的数据速率和 MAE(平均绝对误差，在这里用于衡量估计的姿态参数，如关节点位置、角度等，与真实姿态参数之间的误差)。对于每种算法，数据速率的交叉点位于左侧，而 MAE 的交叉点位于右侧。这表明EventPS实现了更小的MAE和更好的数据效率。

四、展望

本篇的亮点是提出了基于事件相机的立体光学技术，建立了理想朗伯表面产生的事件信号与法向量图的数模，并针对非朗伯面进行优化，最后将事件信号模态融入两个常用的立体光学深度学习模型中。照明的扫描模式有其局限性：“圆形”模式为高仰角表面法线留下了盲区，“次旋线”模式难以机械实现。另外随着照明扫描速度的增加，由于频率响应，事件信号的质量逐渐下降。实现多样化的扫描模式、实现非机械照明设备以及提高高速照明下的事件信号质量值得进一步探索。