A Voxel Graph CNN for Object Classification with Event Cameras

作者：Yong Deng, Hao Chen, Hai Liu, Youfu Li

期刊：CVPR

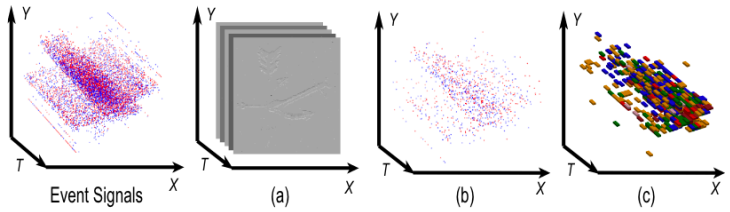
年份：2022

任务：为特定格式的事件数据定制模型以执行核心视觉任务（目标分类）

1、问题背景/研究动机

本文研究的主要问题：**事件相机的目标分类任务**

现有的**面向事件数据的数据驱动方法**分为两类：**基于帧的方法**和**基于点的方法**



- 基于帧的方法(a)

  通过将稀疏事件信号整合到基于帧的**2D**表示中，直接使用**2D CNNs**处理事件数据

  • Pros：利用预训练好的CNN实现更高的性能

  • Cons：牺牲数据的稀疏性，引入冗余信息，需要复杂模型来提取高级特征

- 基于点的方法(b)

  将最初为点云设计的学习模型迁移到事件数据中，**以原始单个事件或区域事件集合作为输入单元**进行特征提取和聚合。

  • Pros：可以事件数据的稀疏性，使用的模型是轻量级的

  • Cons：很难有效地从事件数据中提取区域二维语义，对噪声和场景变化敏感，限制了对复杂场景的泛化能力。

- 逐点输入（以事件点为处理单元）是否适合基于事件的视觉任务?

  • 三维点云中的每个点都可以作为构建物体外部结构的关键点。这些点可以描述几何图形，是对三维物体进行分类的关键。

• 基于事件的分类模型要求能够准确地从事件数据中提取**二维语义**，通常包含**运动轨迹**。

• 作者认为使用原始事件作为输入是不合适的，因为**稀疏的事件点难以为基于事件的模型提供决定性的特征**。

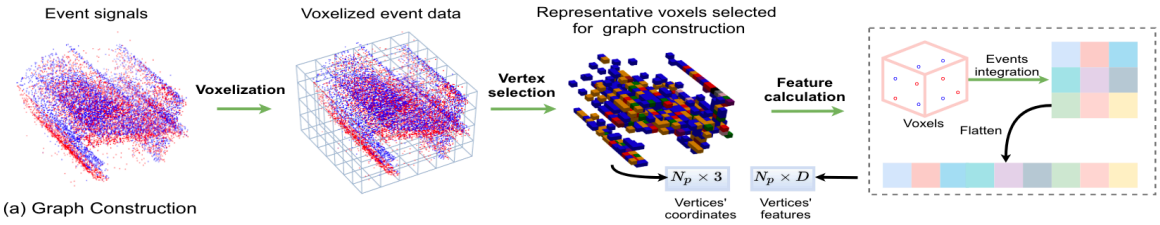
2、解决方法

（1）提出了一种新的**基于事件的图表示构造方法**：有效地利用网格化事件流中的信息特征，在保持稀疏性优势的同时，保留更多的语义和运动信息。

（2）设计了一种轻量级的基于图的学习架构（EV-VGCNN），**引入由多个SFRL组成的MFRL模块**，根据顶点与其相邻点之间的时空关系，从基于事件的图中判别性地学习空间语义和运动信息。

3、具体实现

- 基于事件的图表示方法



• 1. 对事件流进行**网格/体素化**，将每个网格作为图中的一个顶点。

  - 每个顶点的三维坐标由网格的位置（即图的图节点坐标）确定
  - Pros：更好地利用事件数据的稀疏性，并保持数据的稀疏性优势

• 2. 顶点的选择：只保留具有**代表性的顶点**进行图的构建。

  - **代表性的顶点**：网格内部事件点数量最多的Np个顶点
  - Pros：作为噪声过滤器用于输入数据的净化

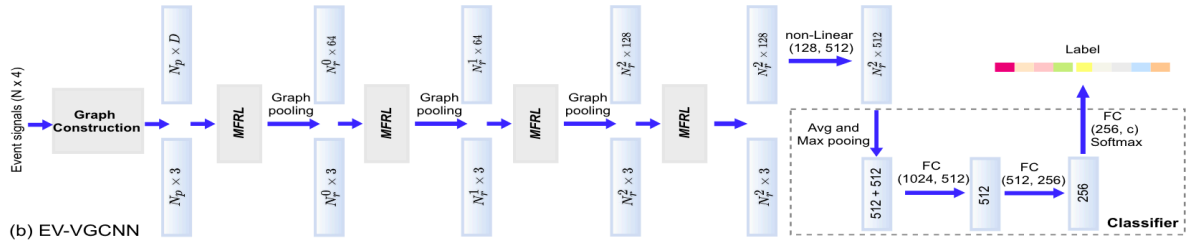
• 3. 顶点的**特征提取**与计算

- 每个顶点的特征由其对应网格内部事件点的特征累积得到（即将网格内部事件点沿时间轴累加到**2D**帧块）
- Pros：通过累积事件点特征，可以提取每个顶点的语义信息。

• 4. 图的构建

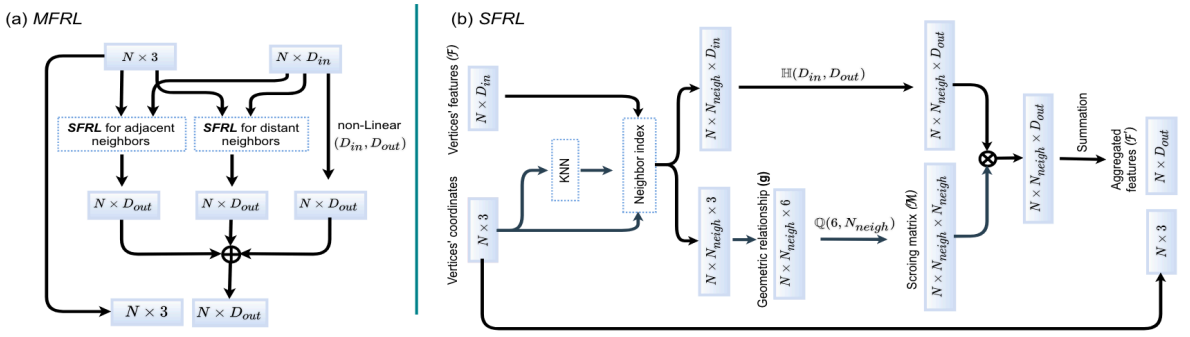
- 采用顶点坐标的距离判断连接关系（隐含了空间/时间上的数据关联，即时空关系）
- 图中每个顶点可以类比为**包含局部纹理和轮廓**等重要线索的静止图像的帧块，有助于网络有效地识别二维场景
- 下游任务完全可以利用任何图网络的知识

- **轻量级基于图的学习架构(EV-VGCNN)**



所提出的学习架构包括3个主要组成部分：多尺度特征关系层（MFRL）、图池化操作和分类器。

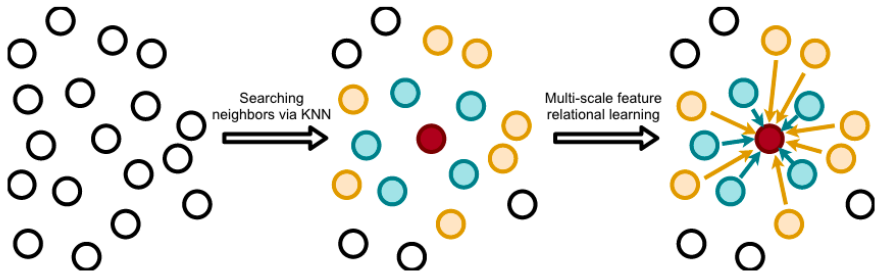
• 多尺度特征关系层（MFRL）



MFRL用于从基于事件的图中的顶点**提取运动和语义信息，**由一个捷径连接和**两个单尺度特征关系层(SFRL)**组成。

- 近相邻点通常携带**局部空间信息**，远相邻点通常包含**运动信息**

- SFRL模块分别负责从近处和远处相邻点中提取特征
- SFRL模块将顶点及其坐标和特征作为输入，并基于每个顶点与其相邻顶点之间的时空关系，计算一个评分矩阵，然后利用该矩阵对相邻点的特征进行聚合。



- 图表示：如何从其近处和远处的相邻点中聚合一个顶点的特征
  - 红点：图中的一个顶点
  - 蓝点：近相邻点
  - 黄点：远相邻的
  - 黄色箭头和蓝色箭头代表从两个SFRL中学习到的权重

- 图池化操作

逐步减少网络中的顶点数量

- 分类器

在高级特征上应用对称函数，以实现对输入的全局表示。具体来说，使用最大池化和平均池化操作分别处理这些高层特征，然后将它们串联起来形成一维特征向量。最后，将全局特征向量送入三个全连接层进行分类。

4、实验对比

文中使用几个基准数据集来评估所提出的方法在**分类精度、模型复杂度**和**浮点运算次数(FLOPs)**上的表现。

- 数据集

5个具有代表性的基于事件的分类数据集：N-MNTIST (N-M)、NCaltech101 (N-Ca1)、CIFAR10-DVS (CIF10) 、N-CARS (N-C)和ASL - DVS (ASL)

• N - M、N - Cal和CIF10是通过记录具有固定运动轨迹的显示器上显示的传统图像获得的，会受到拍摄和仿真环境引入的人工噪声的影响。

• N - C和ASL是使用事件相机在真实环境中记录的，这两个数据集上的评估结果能更好地反映基于事件的模型在实际中的表现。

- 分类准确率

与两种主流的基于事件的分类方法进行比较：基于点的方法和基于帧的方法

• 与基于点的方法的比较

Method	N-M	N-Cal	N-C	CIF10	ASL
H-First [35]	0.712	0.054	0.561	0.077	-
HOTS [26]	0.808	0.21	0.624	0.271	-
HATS [44]	0.991	0.642	0.902	0.524	-
EventNet [43]	0.752	0.425	0.750	0.171	0.949
PointNet++ [47]	0.841	0.503	0.809	0.465	0.947
PointNet++ [47]†	0.955	0.621	0.907	0.533	0.956
RG-CNNs [3]	0.990	0.657	0.914	0.540	0.901
Ours (w/ SFRL)	**0.992**	**0.737**	**0.944**	**0.652**	**0.962**
Ours	**0.994**	**0.748**	**0.953**	**0.670**	**0.983**

- 本文模型的性能优于基于SOTA点的模型。

- 当将输入事件从点表示替换为图表示时，[47]中的方法准确率提高，证明新引入的基于事件的表示的有效性。

- 引入了一个基线模型，将EV - VGCNN中的MFRL替换为SFRL模块。与区分性地学习远近相邻点的特征不同，在SFRL中平等地对待一个顶点的相邻点。可以看出，MFRL的性能都得到了提升，表明所采用的多尺度学习策略通过考虑相邻点之间的时空关系，可以有效地增强特征的判别性。

• 与基于帧的方法的比较

Method	N-M	N-Cal	N-C	CIF10	ASL
Pretrained on ImageNet [9]					
EST [19]	0.991	0.837	0.925	0.749	0.991
M-LSTM [5] †	0.989	0.857	0.957	0.730	0.992
MVF-Net [11]	0.993	0.871	0.968	0.762	0.996
Without pretraining					
EST [19] ‡	**0.990**	**0.753**	0.919	0.634	0.979
M-LSTM [5] ‡	0.986	0.738	0.927	0.631	0.980
MVF-Net [11] ‡	0.981	0.687	0.927	0.599	0.971
AsyNet [30]	-	0.745	**0.944**	**0.663**	-
Ours	**0.994**	**0.748**	**0.953**	**0.670**	**0.983**

- 使用预训练网络后，EST、M - LSTM和MVF - Net的准确率都得到了提升，尤其是在由传统图像转换而来的两个数据集(N-Cal , CIF10)上。因为基于帧的分类模型可以利用在大规模传统图像数据集上预训练的权重。
- 本文方法获得了比大多数从零开始训练的基于帧的方法更好的结果，表明该方法适合从事件数据中提取可区分的表示。

- 模型复杂度

Method	#Params	GFLOPs†	T(CPU)¶	T(GPU)¶
EST [19]	21.38 M	4.28	27.1 ms	6.41 ms
M-LSTM [5]	21.43 M	4.82	34.8 ms	10.89 ms
MVF-Net [11]	33.62 M	5.62	42.5 ms	10.09 ms
AsyNet [30]	3.69 M	0.88	-	-
EventNet [43]	2.81 M	0.91	9.3 ms	3.35 ms
PointNet++ [47]	1.76 M	4.03	174.3 ms	103.85 ms
PointNet++‡ [47]	1.77 M	4.17	178.4 ms	107.97 ms
RG-CNNs [3]	19.46 M	0.79	1236 ms	-
Ours	**0.84 M**	**0.70**	26.1 ms	7.12 ms

- 本文方法在模型和计算复杂度上显示出巨大的优势：
  - 事件的图表示是通过设置一个事件体素而不是单个事件点作为顶点来构建的，便于后续网络在各种场景中学习可区分的特征。
  - 网络中的MFRL模块可以从每个顶点的语义和运动信息中区分出它与相邻点的距离，使其能够在实现准确性的同时构建一个轻量级的网络。
- 在处理数据集N - C中每个样本的平均计算时间上，本文模型的处理速度与基于帧的方法(eg. EST)相同。但与在点网络上的EventNet[43]相比，本文方法在计算速度上表现出弱点。
  - 图表示的集成操作需要耗费大量的计算时间。
  - 相邻点搜索和特征嵌入函数，虽然在很大程度上提高了模型的性能，但也增加了模型的计算时间。

- 消融实验

  - 近远相邻点个数的不同取值比较

	Value						
Variants	A	B	C	D	E	F	G
N_{neigh}^{adj}	10	10	10	5	20	5	20
N_{neigh}^{dis}	15	10	20	15	15	20	5
Accuracy	0.748	0.742	0.751	0.737	0.740	0.743	0.730

- 比较设置A、B、C可以发现，当近相邻点固定时，远相邻点值越大，性能越好。

  - 更远的相邻点来聚合顶点特征时，携带更多的全局信息和时空关系。

- 比较设置A、D、E可以发现当远相邻点固定时，近相邻点的值从10变化到20，最终的性能会显著下降。

  - 只有少量的近相邻点具有表征顶点局部语义的信息。如果相当一部分近相邻点实际上距离较大，那么这些相邻点很难刻画这个顶点的局部语义，容易产生干扰

  - 基于网格的图表示方法的有效性

该部分将本文的基于网格的图与点图进行比较。将这两个图输入到同一个模型EV - VGCNN中，并在N - Cal数据集上测试性能。

点图是通过选择事件数据的随机子集作为顶点，并将事件的极性分配给顶点作为其特征来构建的。

Vertex type	**#Vertex**	**Accuracy**	**GFLOPs**
Original events	2048	0.565	0.63
Original events	4096	0.601	0.66
Original events	8192	0.619	0.72
Event voxels (Ours)	2048	0.748	0.70

- 结果表明，在内部具有相同数量的顶点（2048个）的情况下，基于网格的图表示方法有助于准确率提升。即便将点态图的顶点数增加，本文的方法仍然具有更好的准确率。证明本文方法从事件数据中编码了比逐点图更多的信息特征。

5、局限与启示

- 局限性

• 本研究基于这样一个假设，即输入到模型中的事件数据总是与目标相关。当这一假设不成立时，邻点搜索策略可能无法为顶点找到合适的相邻点。

• 缺乏来自大型数据集的先验知识支持。

• 所采用的同步处理模式通常能够提供比异步方法更鲁棒的全局特征，但不可避免地牺牲了实时性。

- 启发

• 在分类任务中：与传统的三维点云坐标仅表达几何信息不同，事件数据包含两种不同类型的线索，即二维空间信息和运动信息。

• 图表示方法中：网格的筛选与描述以及后续图的连接关系方面可以进一步改进。

• 基于事件的图表示，除目标分类外是否还适用于其他任务。

6、总结

- 提出了一种新的基于图的事件数据学习框架。
- 为了充分利用事件数据的稀疏性并构建低复杂度的模型，提出了一种基于网格的事件数据图表示方法。具体来说，使用网格作为顶点，而不是之前的点作为输入，在保持事件数据稀疏性的同时，也比以往的基于点的方法保留了更多的局部信息。
- 提出了多尺度特征关系层（MFRL），用于从每个顶点的相邻点中区分性地提取空间和运动特征。