学士学位本科生

开题报告

题目：基于动态视觉传感器（DVS 相机）的手势识别系统

学　　号：1609030225

学科专业：软件工程

年　　级：2016 级

姓　　名：王建

指导教师：郭磊

报告时间：

中　国　石　油　大　学

开题报告

1．选题目的及研究意义

1.1 选题目的

神经形态视觉使用硅视网膜传感器采集数据，例如动态视觉传感器（DVS 相机）。动态视觉传感器受到生物视觉的启发，产生的异步事件流指示摄像头视场内各个点的亮度变化。因此，他们能够快速的捕捉变化，天然的消除冗余的背景数据，极大地减少了要处理的数据量，并且它们的动态特性使其非常适合诸如光流，对象跟踪，动作识别或动态场景理解之类的领域。相比传统相机，DVS 相机的具有帧速高的特点，能够达到 20k FPS 的量级，更能够捕捉高速运动和复杂光照条件下的物体。通过动态传感器（DVS 相机）可以实时捕捉突然出现的物体目标，同时能够有效地降低延迟，使得手势识别能够在更多的领域（家电、手机等市场）得以应用。

手势是人类交流和表达的一个自然而直观的部分。我们几乎不需要额外的智能处理就能够使用手势与周边环境中的设备进行通信。在最近几年，深度学习得到了快速发展，并且在机器视觉，自然语言理解等上取得了很好的成效。现有的 AlexNet、GoogLeNet 等算法已经在图像分类上取得了很好的成果，这说明卷积神经网络在机器视觉领域具有良好的应用前景，已成为许多科学研究领域的研究热点之一。

1.2 研究意义

本课题面向的动态视觉传感器在机器视觉领域十分重要。卷积神经网络在图像处理方面的突破近年来吸引了大量研究者关注这一领域。但是对于高速运动的物体，使用传统相机需要以几千或几万帧/秒的速率才能捕捉到，相机处理能力与计算能力的不平衡的瓶颈，限制了其进一步向纵深发展。因此在图像处理方面的优化从而加速机器视觉发展的工作亟待解决。同时，类似的方法同样也可稍微改进以应用于深度学习的其他方面。

动态视觉传感器的速度不受传统的曝光时间和帧速率限制，像素的相应时间在纳秒级；更重要的是，这种传感器可以有效的过滤背景数据，从而成千倍的节省数据量，降低系统成本，并使实时的手势识别变得很容易。

2．国内外研究现状

2.1 动态视觉传感器研究现状

动态视觉传感器（DVS 相机），又称 event-based camera,事件相机，是研究者们受人类视网膜原理启发设计的一种神经拟态视觉传感器。“大脑是具有想象力的，这让我很兴奋;我想制造一种能想象出某些东西的芯片” 这就是 1986 年加州理工学院的一名研究生 Misha Mahowald 开始和 Carver Mead 教授合作从生物学和工程学的角度研究了立体视觉问题的原因。几年后，也就是 1991 年，“Scientific American”杂志封面上的一只猫的照片，展示了一种新的、强大的计算方法，引发了神经形态工程学的新兴领域。这张照片是由模仿人眼神经结构的新型“硅视网膜”拍摄的。

2003 年 Ruedi 等人在空间对比和局部定位方面实现重大突破，同时 Grenet 等人在 2005 年利用该类传感器实现了第一个工业应用。但直到 2006 年，DVS（dynamic vision sensor）的发展才算真正的打开了神经拟态视觉传感器的大门。

动态视觉传感器又称事件相机，是受生物启发的传感器，其工作原理与传统相机截然不同。它们不是以固定的速率拍摄图像，而是异步地测量每个像素的亮度变化。这将引起一系列事件，这些事件将编码亮度变化的时间、位置和符号。事件相机比传统相机相机拥有绝佳的属性:非常高的动态范围(140 db 比 60 db),高瞬时分辨率(达到微妙级别),低功耗,不要受到运动模糊的干扰。因此，在高速、高动态范围等对于传统相机具有挑战性的场景中，动态视觉传感器在计算机视觉方面更具很大的潜力。然而，需要新的方法来处理这些传感器不同于普通相机的输出图像，以释放它们的潜力。

2.2 手势识别研究现状

手势识别是人机交互研究领域的一个重要分支，是机器视觉领域的一个重要研究内容，其发展状况与人机交互技术紧密相连。成熟的手势识别产品已在世界范围内广泛使用，典型的手势识别类电子设备有：2003 年，Sony 公司推出一款名为 EyeToy 的手势识别设备，这种设备能够将玩家的动作传输到游戏画面，实现与玩家的互动；2010 年，微软公司推出的 Kinect 体感设备在手势跟踪与识别方面有着出色的表现，它能实时识别用户手势，结合 Xbox，是用户完成对游戏的控制指令；2012 年，三星推出的智能电视

ES8000 结合用户的手势，可以对电视进行换台、搜台以及音量调节等操作；2016 年，宝马 7 系引入了手势识别功能，驾驶员可以调高或调低音量，接听或拒绝电话等。

近几年的手势识别的研究，大多围绕着传统相机或者深度相机进行。Christian Zimmermann 和 Thomas Brox[1]提出了一个从常规 RGB 图像中估计三维手势的方法，建立了一个深度卷积网络，在此之前学习网络隐式 3D 关节。 连同图像中检测到的关键点一起，该网络可以很好地估计 3D 姿态。Christian Zimmermann 和 Thomas Brox 引入了基于合成手模型的大规模 3D 手姿势数据集，用于训练所涉及的网络。 在包括手语识别在内的各种测试集上进行的实验证明了在单色图像上进行 3D 手势估计的可行性。英伟达公司的 Pavlo Molchanov 等[2]提出了基于循环三维卷积神经网络的动态手势在线检测与分类，实现了实时手势识别，避免了在执行手势及其分类之间出现明显的滞后。手势视频以短片段的形式呈现给 3D-CNN 网络，用于提取局部时空特征。 这些提取出来的时空特征输入到循环网络中，该网络汇总了多个片段之间的过渡，最后通过 softmax 函数进行分类。西安电子科技大学的张亮等[3]通过在卷积网络 LSTM 基础上加上注意力机制，提出了“RES3D+ConvLSTM+Mobilenet“网络结构，并对 ConvLSTM 提出的变体在大规模孤立手势数据集 Jester 和 IsoGD 上进行了评估，取得极佳效果。

基于动态视觉传感器的手势识别研究这几年也越来越多。IBM 研究所的 Arnon Amir 等[4] 提出了首个端到端基于事件的硬件上实现的手势识别系统，该系统使用 TrueNorth 神经突触处理器以低功耗实时地从动态视觉传感器（DVS）流式传输的事件中实时识别手势。Arnon Amir 等首次使用 TrueNorth 处理实时 DVS 事件流，TrueNorth 是具有 100 万个尖峰神经元的基于事件的本地处理器。 TrueNorth 芯片在这里配置为卷积神经网络（CNN），以 105 ms 的延迟识别手势的开始，而消耗的功率却不到 200mW。 CNN 在一个新收集的 DVS 数据集（DvsGesture）上达到了 96.5％的样本外准确度，该数据集包含 3 种光照条件下来自 29 个对象的 11 种手势类别。由于 DVS 具有功耗低、时间分辨率高、动态范围大、存储需求少等优点，Stefanie Anna Baby 等[5]探讨了使用 DVS 进行人类活动识别（HAR）的可行性。他们提出使用 DVS 视频的各种切片作为 HAR 的特征映射，并将它们表示为运动图。他们发现将运动地图与运动边界直方图（MBH）融合，在基准 DVS 数据集以及我们收集的真实 DVS 手势数据集上具有良好的性能。

2.3 发展趋势

机器视觉已经越来越多地代替人类进行视觉信息的获取，帮助处理复杂的现实生活环境。靠着算法的提升，机器视觉的处理速度、跟踪物体准确能力已经有了很大的改善，但是在处理复杂的现实环境时，还是会有误差大、计算量大、功耗大的问题。模仿人类脑神经元的卷积神经网络打造出了强大的人工智能，模仿人类视网膜结构也应当能给机器配备强大的视觉能力。

在动态视觉传感器中，对于单个像素点，只有接受光强度发生改变时，才会有事件(脉冲)信号输出。DVS 可以拍摄高速运动的物体，即便物体运动非常快，也能拍摄清楚并且以任何速度播放。而且它比高帧率摄像机更能胜任这项任务，不会如同基于帧率的摄像机那样产生大量的冗余数据，对运算资源、能耗造成浪费。

与基于帧的摄像机相比，基于事件的动态视觉传感器通过在像素检测到亮度变化时发送异步事件来模仿生物视网膜，从而消除了多余的数据传输。用户做出的手势事件经过动态视觉传感器处理后传递到卷积神经网络进行手势识别。

3. 相关技术原理和技术路线

首先利用动态视觉传感器收集 DVS 数据集，完成数据的成帧处理。对于静态的手势识别，处理得到的数据集是各个手势（例如剪刀、石头、布）的单个图片，将数据传入 AlexNet 网络[7]进行训练；对于动态的手势识别，处理得到的数据集是对应动作（例如左挥手、右挥手、抖手等）连续的图片或视频片段，将数据传入由 3DCNN 和 LSTM 所搭建的网络进行训练。

3.1 . 技术原理

3.1.1 静态手势识别神经网络训练算法（AlexNet）

Alex 在 2012 年提出的 alexnet 网络结构模型引爆了神经网络的应用热潮，并赢得了 2012 届图像识别大赛的冠军，使得 CNN 成为在图像分类上的核心算法模型。AlexNet 该模型一共分为八层，5 个卷积层,，以及 3 个全连接层,在每一个卷积层中包含了激励函数 RELU 以及局部响应归一化（LRN）处理，然后在经过降采样（pool 处理）。

输入	227*227*3 （RGB图像，3通道）		卷积层3	输入尺寸（Conv3_In）	13*13*256
卷积层1	输入尺寸（Conv2_In）	227*227*3		卷积核大小（kernel_size）	3*3
	卷积核大小（kernel_size）	11*11		卷积核个数（kernel_num）	384
	卷积核个数（kernel_num）	96		滑动步长（stride）	1
	滑动步长（stride）	4		边界填充0的个数（padding）	1
	边界填充0的个数（padding）	0		输出尺寸（Conv3_Out）	13*13*384
	输出尺寸（Conv1_Out）	55*55*96	卷积层4	输入尺寸（Conv4_In）	13*13*384
池化层1	池化单元大小（kernel_size）	3*3		卷积核大小（kernel_size）	3*3
	滑动步长（stride）	2		卷积核个数（kernel_num）	384
	池化单元个数（kernel_num）	27		滑动步长（stride）	1
	输出尺寸（Pool1_Out）	27*27*96		边界填充0的个数（padding）	1
卷积层2	输入尺寸（Conv2_In）	27*27*96		输出尺寸（Conv1_Out）	13*13*384
	卷积核大小（kernel_size）	5*5	卷积层5	输入尺寸（Conv5_In）	13*13*384
	卷积核个数（kernel_num）	256		卷积核大小（kernel_size）	3*3
	滑动步长（stride）	1		卷积核个数（kernel_num）	256
	边界填充0的个数（padding）	2		滑动步长（stride）	1
	输出尺寸（Conv2_Out）	27*27*256		边界填充0的个数（padding）	1
池化层2	池化单元大小（kernel_size）	27*27*256		输出尺寸（Conv5_Out）	13*13*256
	滑动步长（stride）	2	池化层5	池化单元大小（kernel_size）	3*3
	池化单元个数（kernel_num）	13		滑动步长（stride）	2
	输出尺寸（Pool2_Out）	13*13*256		池化单元个数（kernel_num）	6
				输出尺寸（Pool5_Out）	6*6*256

Figure 1 AlexNet **卷积层**

全连接层6	全连接层输入	6*6*256
	全连接层输出	1024
	dropout	
全连接层7	全连接层输入	1024
	全连接层输出	1024
	dropout	1024
全连接层8	全连接层输入	1024
	全连接层输出	5

Figure 2 AlexNet **全连接层**

通过 AlexNet 网络对物种静态手势（剪刀、石头、布、OK 和点赞）进行训练，最后输出的分类结果只有五种，需要对 AlexNet 网络的全连接层进行调整，将全连接层的 4096 个过滤器减少为 1024 个过滤器，最后输出 5 种分类结果。

3.1.2 动态手势识别神经网络训练算法

3D-CNN[8]能够直接从原始输入中提取特征，通过执行 3D 卷积在视频中从时间和空间维度提取特征，将高级功能模型规范化，并结合各种不同模型的输出，进一步提高

3D CNN 的性能。3D-CNN 在视频分类，动作识别等领域发挥着巨大的优势。3D 卷积是通过堆叠多个连续的帧组成一个立方体，然后在立方体中运用 3D 卷积核。通过这种结构，卷积层中的特征图都会与上一层中的多个相邻帧相连，从而捕获运动信息。

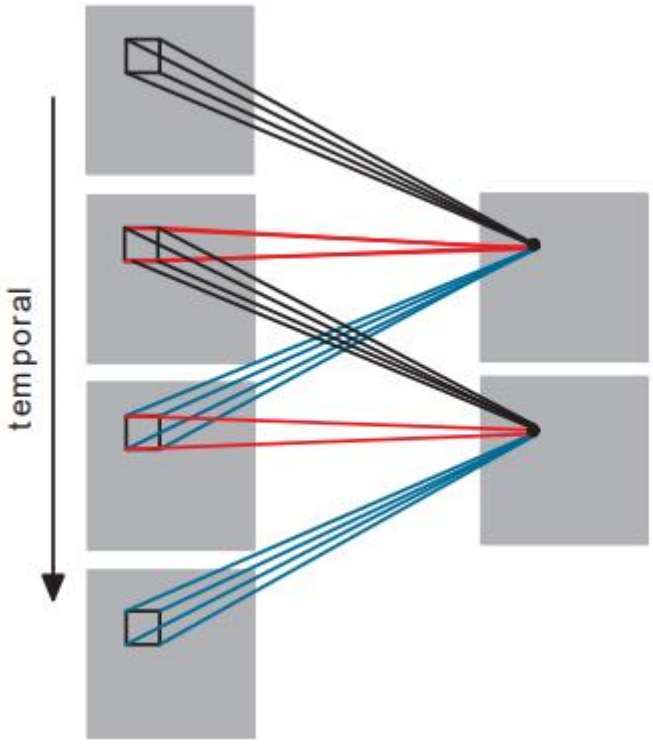

Figure 3 3D-CNN

长短期记忆（Long short-term memory, LSTM）是一种特殊的 RNN，主要是为了解决长序列训练过程中的梯度消失和梯度爆炸问题。简单来说，就是相比普通的 RNN，LSTM 能够在更长的序列中有更好的表现。ConvLSTM 核心本质还是和 LSTM 一样，将上一层的输出作下一层的输入。不同的地方在于加上卷积操作之后，为不仅能够得到时序关系，还能够像卷积层一样提取特征，提取空间特征。这样就能够得到时空特征。并且将状态与状态之间的切换也换成了卷积计算。

3.2 技术路线

基于动态视觉传感器（DVS 相机）的手势识别系统的体系架构设计流程如图。首先解决 DVS 数据处理，包括成帧、去噪等步骤，获得网络训练所需要的数据集。然后分别通过 AlexNet 网络和“3D-CNN+LSTM”网络进行手势识别。

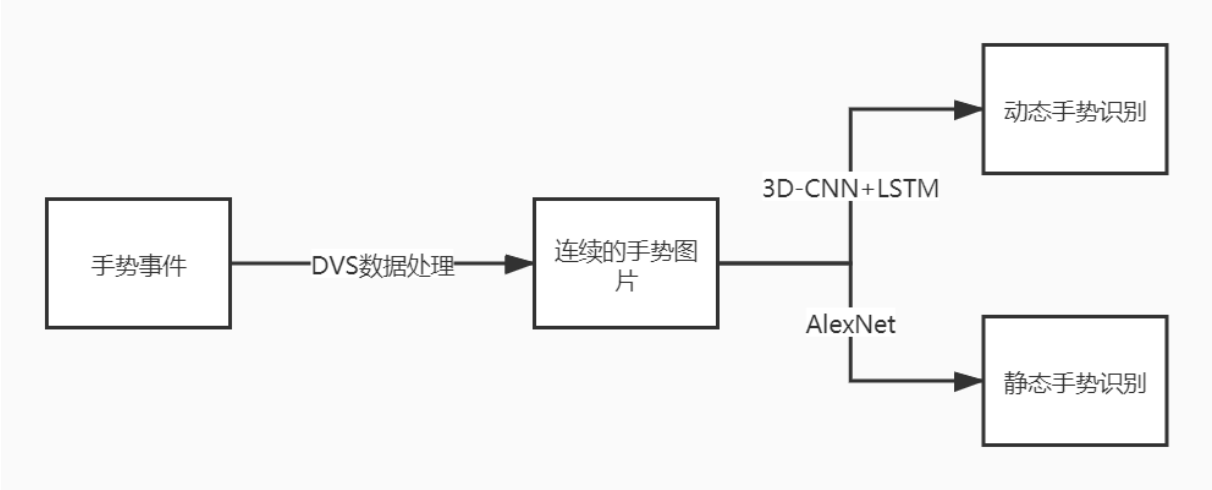

Figure 4 **技术路线图**

4. 研究难点及目前存在的问题

4.1 研究难点

1. DVS 数据的处理。DVS 传送的事件需要经过去噪、渲染、成帧等处理才能转换成神经网络训练所用的数据类型。

2. 实时手势识别网络的搭建。如何将 3D-CNN 与 LSTM 相结合，网络参数的确定。

3. 降低在线的动态手势识别的延迟。对于动态手势识别，需要尽可能减小做出手势与调用模型识别之间的延迟，以实现实时效果。

4. 将实时手势识别和基于事件的手势识别结合。虽然实时手势识别和基于事件的手势识别的研究很多，但是二者结合的相关研究很少。

4.2 存在的问题

在实际的研究和设计过程中，存在如下几个问题：

1. DVS 数据的成帧和去噪处理；

2. 动态手势识别网络的优化；
3. 调用模型识别手势的延迟。

5. 研究预期成果

1. 基于动态视觉传感器（DVS 相机）的实时数据显示。通过动态视觉传感器实时捕捉突然出现的人物目标，实现实时数据显示，同时利用动态视觉传感器制作手势图数据集，并开源数据集；

2. 搭建完整的基于动态视觉传感器的（DVS 相机）的手势识别系统，实现静态的手势分类和动态的手势识别；

3. 在完成毕业论文的同时，发表一篇期刊论文。

6. 进度安排

1. 2020.01.10-2020.03.04　查阅文献，提交开题报告，确定文章的大体结构；

2. 2020.03.04-2020.03.30　完成论文完整提纲，进一步收集相关资料及文献，同时完成 DVS 数据处理和手势识别相关工作；

3. 2020.04.01-2020.05.10　按照毕业设计任务书要求开展研究工作，完成论文的初稿，并与导师商讨不合理之处；

4. 2020.05.10-2020.05.30　修改论文并定稿，提交导师检查；

5. 2020.06.01-2020.06.15　排版、打印装订，上交论文，准备答辩。

参考文献

[1] ChristianZimmermann, Thomas Brox. Learning to Estimate 3D Hand Pose from Single RGB Image[J]. 2017:4913-4921.

[2] Molchanov P , Yang X , Gupta S , et al. Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks[C] 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[3] Zhu Guangming,Zhang Liang,Yang Lu,Mei Lin,Shah Syed Afaq Ali,Bennamoun Mohammed,Shen Peiyi. Redundancy and Attention in Convolutional LSTM for Gesture Recognition.[J]. IEEE transactions on neural networks and learning systems,2019.

[4] Liang Zhang, Guangming Zhu, Lin Mei, Peiyi Shen, Syed Afaq Ali Shah. Attention in Convolutional LSTM for Gesture Recognition[C]. NIPS,2018.

[5] Amir A , Taba B , Berg D , et al. A Low Power, Fully Event-Based Gesture Recognition System[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

[6] Stefanie Anna Baby, Bimal Vinod, Chaitanya Chinni, Kaushik Mitra. (2018). Dynamic Vision Sensors for Human Activity Recognition. 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)

[7] Krizhevsky A , Sutskever I , Hinton G . ImageNet Classification with Deep Convolutional Neural Networks[C]. NIPS, Curran Associates Inc, 2012.

[8] Ji, Shuiwang et al. “3D Convolutional Neural Networks for Human Action Recognition.” IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (2010): 221-231.