**基于嵌入系统与深度学习的**

**人体姿态和生理状态视频监测系统**

**开题报告**

专业综合实践训练2021

生医8 李浩伟 李子涵 江柔蓝

目录

[一、项目背景 2](#_Toc70088884)

[二、需求分析 2](#_Toc70088885)

[三、技术路线 3](#_Toc70088886)

[2.1 基于Horizon X3开发板的视频检测处理技术 4](#_Toc70088887)

[2.2 基于深度学习的人体姿态分类： 6](#_Toc70088888)

[2.3 基于视频的心率检测算法 9](#_Toc70088889)

[五、实用性、可行性与创新性分析 10](#_Toc70088890)

[六、工作进度安排 10](#_Toc70088891)

[七、小组分工计划 10](#_Toc70088892)

[八、经费预算说明 11](#_Toc70088893)

[九、参考文献 11](#_Toc70088894)

# 一、项目背景

人口老龄化是世界性问题，在中国的发展尤为迅速。国家统计局发布的《中华人民共和国2019年国民经济和社会发展统计公报》显示，中国60周岁及以上的老年人口数量已达到2.5亿，占人口总数的18.1%，其中65周岁以上占人口总数12.6%[1]，这些数据说明中国已经进入老龄化社会。社会各界都十分重视养老问题，给健康监护及远程医疗领域带来了更高的机遇和挑战。

家庭养老是当前养老模式的主流之一。在中国特有的“4+2+1”家庭结构下，子女多因工作无暇照顾老人，老人独居在家中通常面临着一系列的问题，一方面其在生活中更容易出现突发的危险，如摔倒等，而这些事件的发生往往会对老龄人的身体健康产生很严重的影响，或者预示着老年人的健康状况已经出现了重要变化；另一方面，老年人的日常作息，如一天中坐下、站立、躺卧、行走时间会对老年人的身体健康产生很重要的影响，也能反映老年人的健康状况，这些统计数据将为“大健康”战略的发展提供支撑。

因此，我们希望设计开发一个基于视频的多功能健康状态监测系统，重点实现对老年人的行为模式的实时识别、统计和反馈，为家庭养老的安全和健康提供助力。

近年来，家用监控设备已经被相当多的家庭所接受，全球著名消费科技市场研究机构Strategy Analytics发布最新的研究报告《2019年智能家居监控摄像头市场预测和分析》指出，2019年全球消费者将在智能家居摄像头上支出近80亿美元；在美国、英国、德国和法国中，有将近三分之一的家庭安装了监控摄像头。在中国，选择安装家用监控摄像头的家庭也已经超过10%，并且正在迅速增长。由此我们认为，基于视频的健康检测系统是能够被大部分家庭所接受的。

为了实现对人体姿态的检测识别，目前较为常见的系统可以分为基于视频图像和基于可穿戴式传感器两类[2][3]。与基于可穿戴式传感器的姿态检测系统相比，基于视频的方法具有“非接触”的优点，安装方便，能够提供长时间、稳定的检测，有效避免传感器丢失、损坏、电量不足等情况，且不会对老人的日常生活造成不便。另一方面，基于视频的方法可以一次性采集多种参数，相当于在被观测者的不同部位同时佩戴了多组传感设备，且具有很强的可拓展性，可以迅速将计算机视觉领域的新兴研究应用在该系统中。此外，在“居家养老”这一特定的应用场景下，室内环境和周围光线的变化较为简单，非常适合视频检测系统的使用。

# 二、需求分析

**总目标：**设计开发一个基于视频的多功能健康状态监测系统，重点实现对老年人的行为模式的实时识别、统计和反馈，为家庭养老的安全和健康提供助力。

**阶段目标与关键问题：**

通过现有开发板，提取视频中人物的骨架信息。后续算法均需部署在开发板上。在简单室内环境和单一目标下，通过深度学习实现基于骨架信息的实时坐、站立、躺卧、行走等姿态识别功能。给出相应的统计信息。实现基于视频的心率检测。

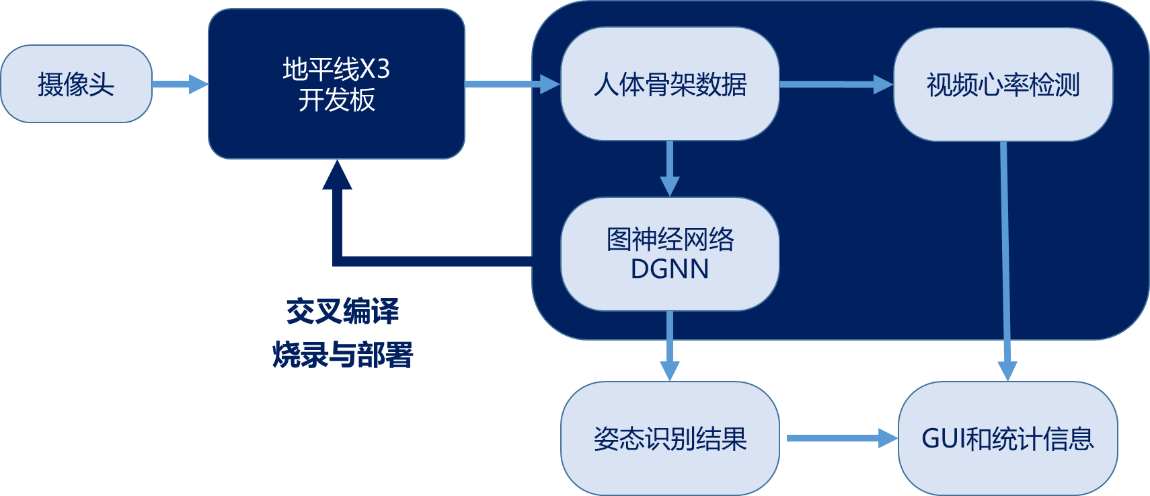
**提高目标：**多目标的识别和检测问题，尝试更多姿态识别（如吃饭、吃药等等），处理较复杂环境如光线变化等。

# 三、技术路线

项目的核心技术是人体行为模式的提取。对于这个目标现存许多不同的解决方案。包括基于各种传感器的检测和基于视频的检测等等。视频检测因为不需要传感器佩戴，方便子女进行实时的照看等等特性具有较好的探索潜力。普遍得讲，行为模式得识别一般分为人体追踪和模式识别两个部分。在前一部分即存在光流法等纯视频分析方法，也有open pose，yolo，RCNN等深度学习方法。而模式识别较为广泛的方式为先进行骨架或特征点的提取，而后通过特征点的空间信息进行识别。NTU-RGBD数据集，作为一个广泛使用的行为识别的数据集，在近些年已经验证了基于骨架图的模式识别在正确率上可以达到一个非常高的水平，这些网络通常都是基于图网络的结构。但是这些都是对于本地视频的行为检测，对于实时的检测，问题则困难的多。近些年有人提出一些轻量级网络，实现了高达180fps的检测，虽然这些网络往往以一些准确度的牺牲为代价，但是其预示着附带端处理功能的小型化的家用行为检测视频设备是可能的。

地平线作为一家国产公司，在计算机视觉方面有着出色的成就，和众多的智能汽车厂商保持着紧密的合作，提供智能驾驶的解决方案。其提供的开发板X3在2.5W的功耗下达到了2TFLOPs/W的高计算能耗比（现有设备中仅有haswell-eq 2.13，Raspberry 2.02）比其略微优秀一些。但是X3在视频编解码速度上占有一定的优势，同时其作为一个放释放的产品，也相应具有更大的发展潜能。因此我们选择该设备作为我们的硬件基础。

脉搏波信号是现在非常常用的获取心率数据的手段。投射式和反射式的光电容积脉搏波测量方法被广泛应用在接触式的心率检测手段中。这种光电容积法本质上是通过主动光的方式获得血管容积变化的信息。但是在现实中，我们并不需要主动光，而在环境中天然得存在我们需要的颜色的光线。因此我们可以通过捕捉人脸对一定波长光线的反射信号捕捉心率。这个过程中诞生了非常多的方法，包括绿色光滤波，HSV空间检测，独立成分分析等等。但是我们可以发现现有的大量论文都是基于静态、稳定、补光的条件下进行的。实际上，人脸上的反射光的变化主要由三个变化导致：环境光的变化，人脸对环境光的反射变化和脉搏波引导的光吸收变化。所以有研究者希望对人脸的三个不同的独立部位进行捕捉，通过独立成分分析的方式对脉搏波信号进行分离。同时，在现有研究中，为了捕捉频率尽量高，一般对于人脸不同部分的识别仅采用haar cascade的方式，其在一些时候会产生较大的偏差。而这个问题恰恰好在上面一部分得到了解决：X3的开发板提供了非常好的算力支撑可以告诉获取部位信息，同时可以告诉对视频信号进行处理。我们希望在这个基础上使用，并尽量优化前人已有的检测算法。同时此处，DEAP数据集提供了大量的视频和心电数据，我们正在争取获得该数据集，以对算法的结果进行更好的验证。



## 2.1 基于Horizon X3开发板的视频检测处理技术

本项目采取地平线2020年新推出的Horizon X3开发板，其硬件内含四核ARM [Cortex-A53@1.5GHz](mailto:Cortex-A53@1.5GHz)，[双核BPU@1.2GHz](mailto:双核BPU@1.2GHz) 拥有5TFLOPS等效算力，对于视频的处理具有相对足够的算力支持。BPU是Horizon公司自主研发的高效人工智能处理器架构，在早期的产品中已经验证了在道路环境下的视频语义处理有非常好的效果。同时开发版内含2GB 三星 LPDDR4，8G eMMC5.1和Hdmi, TF/SD, MIPI CSI, MIPI DSI, UART, 40 pin等全套齐全的接口，为视频留的输入和输出提供了一定的可能。

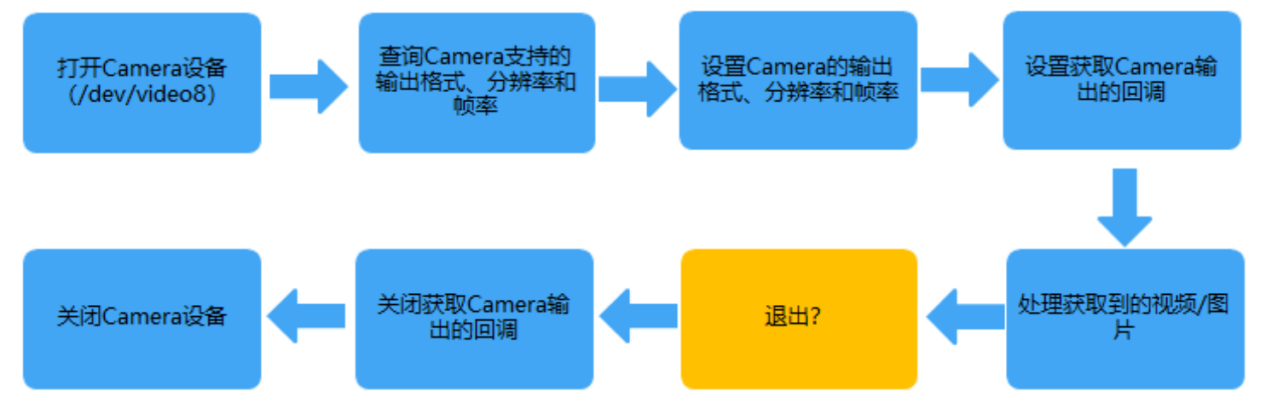
在开发中，我们具体需要的硬件和软件需求如下：

* Linux开发机或虚拟机，Ubuntu或Centos，用来进行AIExpress编译和调试开发
* Windows PC，调试串口，烧录程序和系统
* X3sdb开发板
* 显卡服务器，用以进行模型训练
* 灰点高速相机(1.3MP, 149fps)

开发流程具体如下：

1. 开发板系统烧录，进行基本的系统配置
2. Linux中编写程序，进行调试、利用提供的开发板虚拟软件进行模拟运行。
3. 利用神经网络训练模型，并进行测试（与1同步）
4. 编写基于人脸视频的心率检测代码，并进行测试（与1同步）
5. 软件编译
6. 通过烧录软件或者scp协议把编译好的结果部署在开发板上
7. 运行，并通过网口将数据传输到web端进行展示

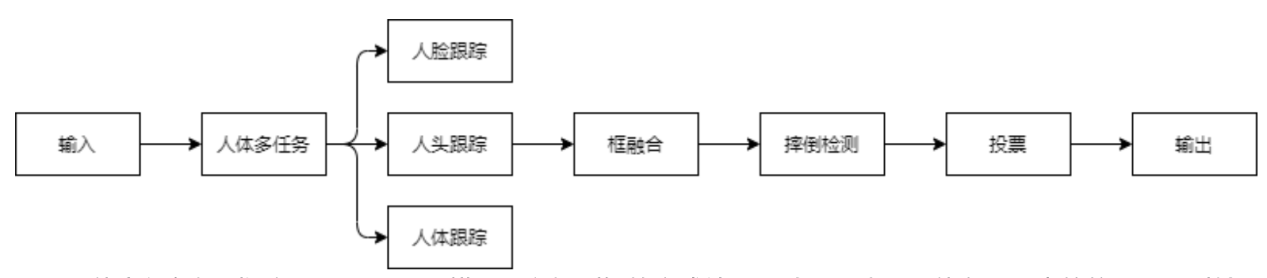
对于一个基于视频流的程序而言，地平线的官方论坛中给出了相应的执行顺序的介绍：



(图源地平线AI论坛, zhuk, <https://developer.horizon.ai/forum/id=5f312d96cc8b1e59c8581511)>

在项目开发板的选择上，我们实际上进行了大量的考量。相对而言，树莓派更类似于一个微型的电脑，其对各种不同的开发需求的支持较好，也就代表了对于某种特定的任务的支持不够充分。不得不承认树莓派等开发版在同样的价格下具有量化指标更高的算力，对于开发也更为友好，但是Horizon X3采取FPGA的方式，在硬件级别对卷积运算等进行了大量的优化，也就导致了其对视频任务有更好的适配。但是这样的硬件的使用也会带来一定的麻烦。由于传输缓冲区换个速度的限制，浮点运算对传输带宽要求更高，因此变相降低了算力的使用。因此开发板采用的是定点数的计算模式，这就要求对模型进行一定的转换。这些转换和编译的过程可能给项目的开发带来一定的困难。

为了一部分克服这些困难，优化开发流程，Horizon公司提供了Ai Express等一系列的工具，并且实时的目标检测和人体骨架点提取的接口。其已经提供的一套摔倒检测的算法流程具体如下：



(图源地平线AI论坛, asdf, <https://developer.horizon.ai/forum/id=5efab48f38ca27ba028078dd>)

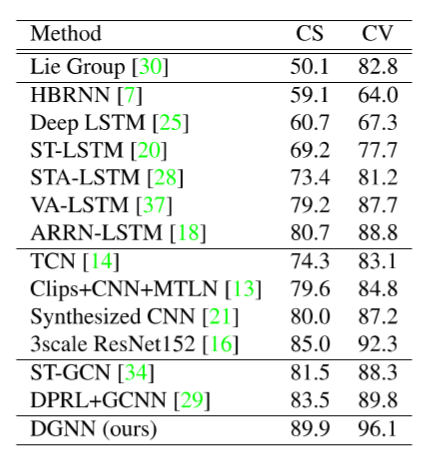
**技术难点**

* 开发板推出时间晚(2020/9/9)，开发社区建构不完善，开发工具集少，开发手段较为局限，在安装调试的时候没有好的的经验或者案例可以借鉴，需要更多得自己探索或和官方开发人员进行沟通
* 开发板现有得硬件支持技术并不成熟，相比而言，其提供了较高得算力，为轻量级的模式识别提供了可能性，但是另一方面其支持的数据格式，硬件种类都 较少。
* 需要交叉编译，多平台联合开发

## 2.2 基于深度学习的人体姿态分类：

由于我们从开发板上获得的数据直接就是人体的关节、骨头信息，因此使用图神经网络来建模这一问题是相当直观的，并且从SOTA的角度，图神经网络也应该是合理的。我们对于图的定义如图1所示。我们定义关节为图的节点，骨头为图的边。边是有向的，我们定义人体胸部正中为根节点，那么边的方向总是由距离根节点更近的节点指向距离根节点比较远的节点。由于边的指向性，我们可以进一步定义指向一个节点的边为，而离开一个节点的边为，并定义，分别为指向顶点vi的边集合与指出顶点vi的边集合。

表1 2020年NTU数据集上各种深度学习算法的正确率



表注：CS是以受试者作为训练集测试集划分标准；CV是以视角作为训练集测试集划分标准。DGNN算法正确率对于我们这个问题应该是合适的。

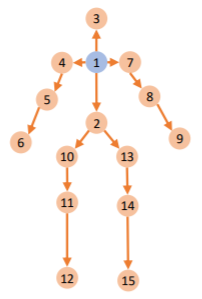
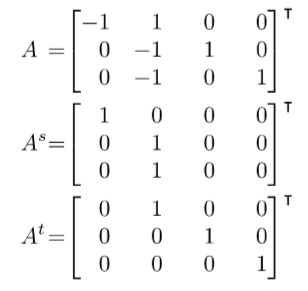


图1 人体骨架的图建模方法

在以上抽象的描述之后我们需要具体化这一张图，假设一张图上有Nv个节点，Ne条边，那么使用邻接矩阵A可以表示为Aij，i=1…Nv, j=1…Ne，如果节点vi是节点边ej的起点，则Aij记为1，终点则记为-1，否则记为0。定义矩阵As=A>0, At=A<0为了后续运算方便。



在充分定义了这样一张图之后，我们希望利用这些图信息来完成动作的分类任务。这就引出了使用directed graph neural network（DGNN）这种网络架构。首先我们需要引入DGN block，它包含了两个更新函数与，同时包含两个聚合函数与，论文给出的聚合方法是平均池化。基于这两个函数以及原有图信息的更新方法如下四个式子所示。其中对于符号需要额外说明的是，第三、四个式子的左值代表的是经过一次图更新之后的节点和边，用“‘”来表示；而他们的右值中[·]代表向量的合并运算。这四个式子以顺序的形式执行，即对于所有节点，首先通过指向这些节点的边的特征来计算出该节点的第一个特征，再通过指出这些节点的边的特征来计算出该节点的第一个特征（式子1，2）；接着，使用该节点的两个特征与以及该节点的原始特征共同计算该节点更新后特征。在对于所有的节点都进行了一次更新之后，就可以利用所有更新后的节点的特征以及边在上一次执行算法后的特征，共同更新出新的边特征。

进一步的，可以将上述的图的矩阵表示与特征推导过程相结合，当我们的聚合函数取到的是平均池化时，在计算过程中其相当于一个单位矩阵（忽略常数项），因此不会在公式中显式标出，而节点和边的特征向量可以用矩阵来表示，即我们定义是一个CT×Nv的矩阵，是一个CT×Ne的矩阵，这里CT是特征的维度数。那么式（1）~（4）可以表示为式（5）~（6）。然而，此时模型仍然是线性的，在此基础上进一步加入激活函数和Batch normalization可以提供非线性以及防止梯度消失或者梯度爆炸。

经过以上的建模过程，我们已经可以很好的对于一帧的信息进行编码。为了进一步利用时间上的信息，我们可以在时间上对于每一个节点与边的特征向量进行一维卷积。最后再通过一个全连接层，即可完成分类任务。同时，为了充分使用时间顺序信息，我们可以进一步的构造特征向量我们让这两个特征向量上文中的节点和边的更新操作进入另一个相同架构的网络，最后也可以通过全连接层输出概率。最后，基于模型融合的思想，将两个模型的概率求和，再以最大的概率输出类标签即作为输出值。

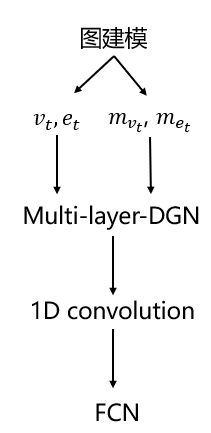


图2 图卷积网络流程图

**需要解决的潜在问题：**

我们同样发现上述的过程在实际实现的过程中会遇到很大的问题。一是NTU数据集对于提取的骨架关键点与我们相机的骨架关键点提取结果不一样。如图3所示，左图是NTU数据集的骨架提取结果，而右图是我们相机的骨架提取结果。可以看出有以下四点比较显著的差异，一是我们相机在肢体只有12个点，而NTU数据集上也有24个点；第二，我们相机时常会出现漏点的情况，比如在图2中，我们的相机就漏掉了左手手腕的点；然而开源的数据集上不存在漏点的可能性；第三，在骨架的提取上，开源数据集的背部是一根线，而我们提取到的数据则是一个四边形；第四，开源数据集中坐标实际是三维的（画图时投影到了二维），然而我们的相机采集的数据却只有二维。基于以上的多种差异，我们计划要首先对开源数据集的点进行采样，更进一步说是将开源数据集适配到我们的采样点上，这样才能在开源数据集上进行大规模的预训练。基于以上的大量问题，目前与王老师组的协商结果是，最终在站立、坐下、躺下、走路这几个类别的识别正确率有70%就相当优秀了

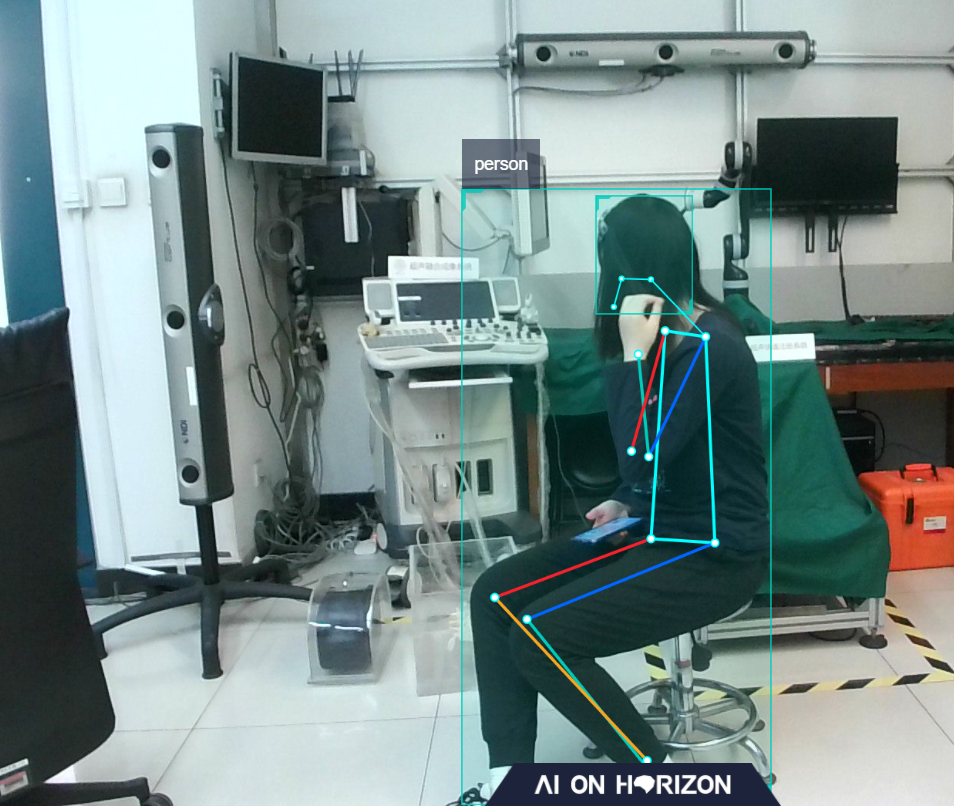
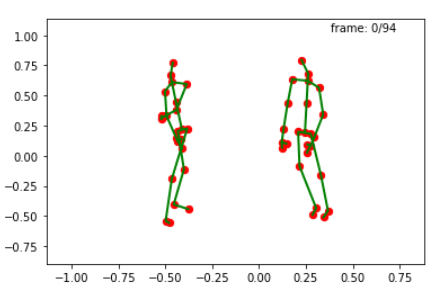


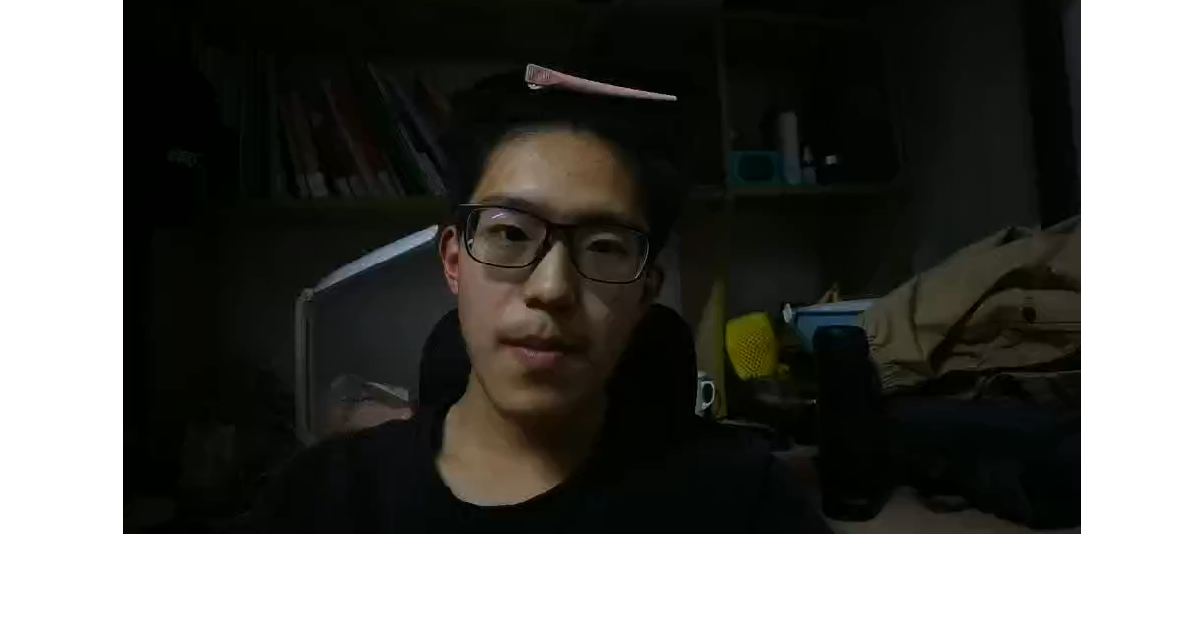
图3 骨架提取结果对比

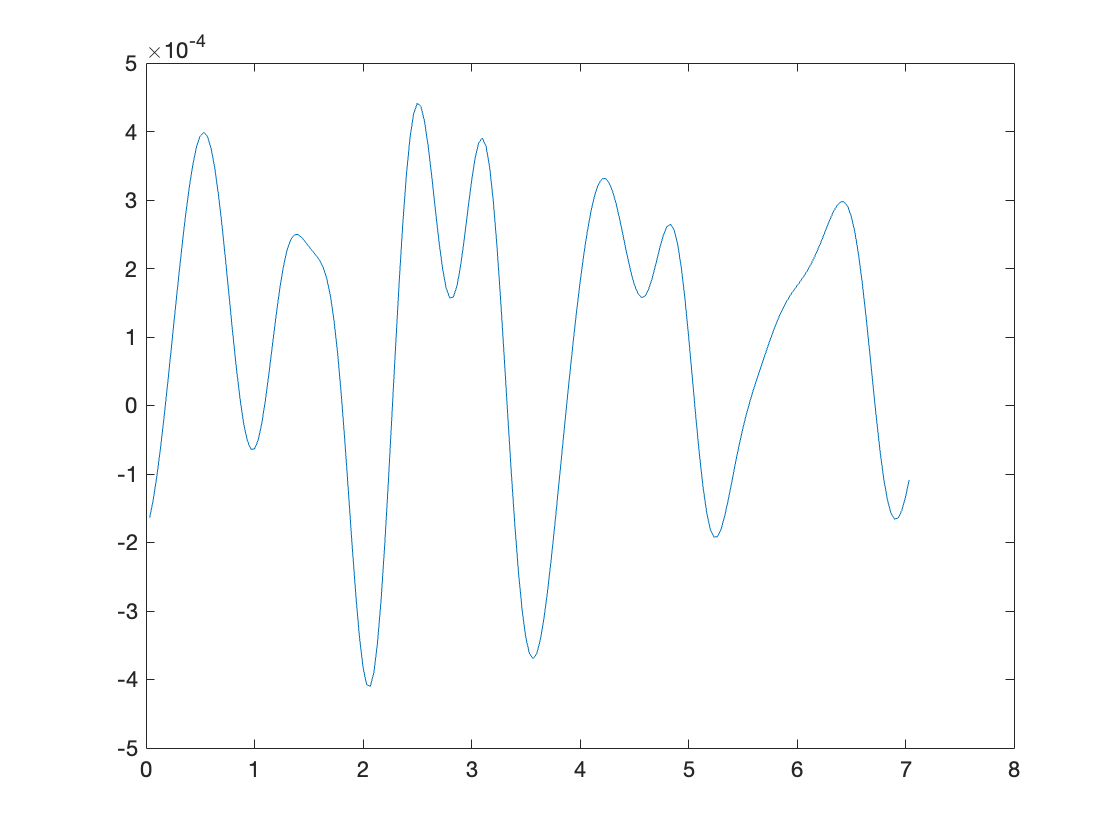
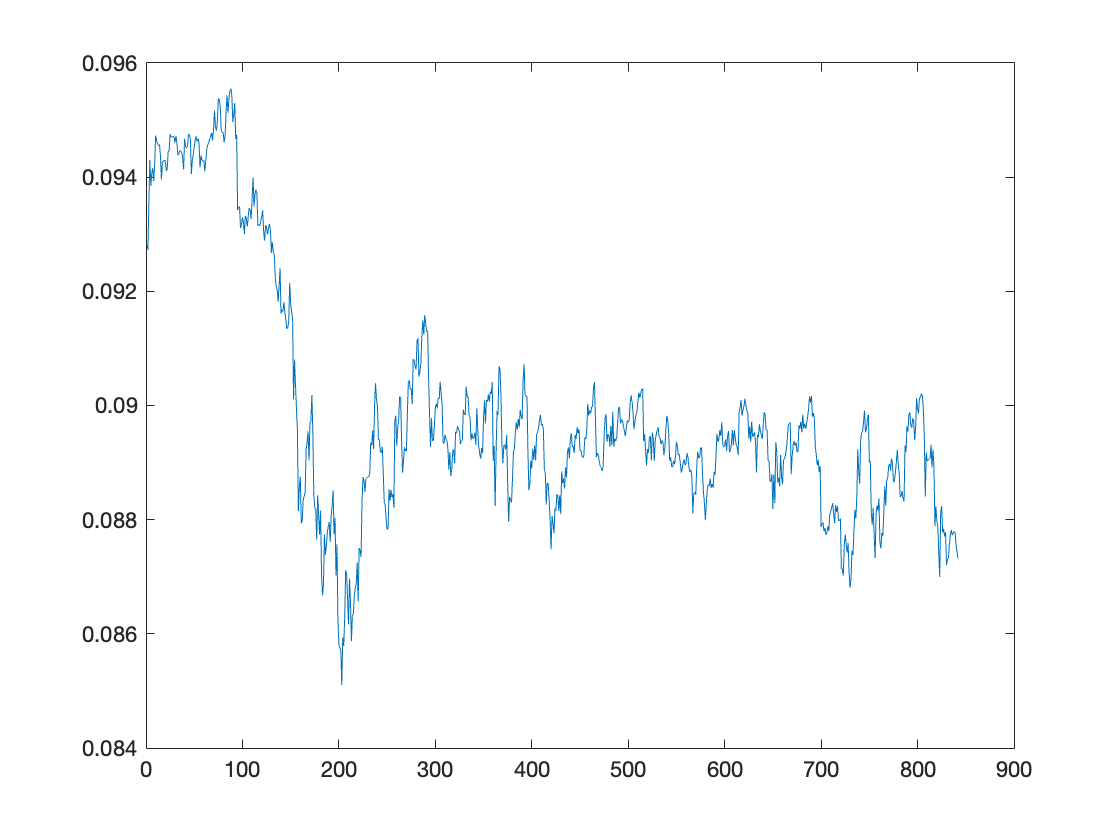
图注：左图时NTU数据集上的骨架提取结果，右图是我们开发板的骨架提取结果

## 2.3 基于视频的心率检测算法

在PPG的使用中，我们使用主动光源，通过测量反射光的信号的强弱来检测脉搏波的变化。但是在正常的环境光场中，如果我们利用相机检测的反射光，理应也能检测到血流变化引导的反射光节律性变化。对于这种检测方案有多种实现路线。最基础的，我们可以通过绿色滤光或者检测H色域波动来探测血流的变化[4]。也有论文指出，我们可以认为血流引导的反射光变化，环境光变化和人体的位移产生的光场分布变化三个共同导致了检测的结果，而且可以认为这三个在数据分布上是独立的，这个时候采用ICA的方法，就可以检测出更稳定的心率信号值[5]。

下面是基于HSV和RGB检测的已经实现的效果图：





我们将继续试探其余的心率检测算法，并在DEAP数据集上进行验证。

在本项目中该流程需要处理的事情（后几个在有时间的情况下完成）

1. 获得每一帧图片的时间信息，帧内的骨架信息，帧内识别人脸部分对应的rgb的图像
2. 实现更多的基本分类，对视频中的多个骨架实现行为模式识别
3. 通过面部图像进行颜色特征提取，通过一些基本的生理学假设、建模和滤波进行PPG的分离，实现心率检测
4. 基于视频进行人脸识别，达成行为模式统计的作用
5. 基于上述的统计结果设计一定的交互
6. 通过多个摄像头的骨架进行三维重建，进行更精确的复原

# 五、实用性、可行性与创新性分析

可行性：本项目组同学具有良好的生物医学工程基础知识，掌握C++/Python/MATLAB等多种编程语言，在硬件调试、计算机视觉、软件开发等领域具有一定的经验。本项目受到王广志老师课题组大力支持，课题组已提供项目核心硬件（地平线X3开发板）和相关技术文档。自第三周起，指导老师已经与项目组同学进行了多次交流。

创新性：将嵌入式系统与深度学习方法相结合，应用在人体姿态识别领域，并针对“智慧养老”这一具体场景给出系统设计。

实用性：老龄化社会中的养老问题正在被越来越多的人所关注，“4+2+1”家庭结构下子女很可能无暇照顾老人，老人独居在家中极易发生危险，其不同姿态的统计数据也与老人的健康息息相关。我们设计开发的基于视频的多功能健康状态监测系统，能够实现对老年人的行为模式的实时识别、统计和反馈，为家庭养老的安全和健康提供助力。

# 六、工作进度安排

第九周前：文献调研，确定技术路线。软件系统准备，服务器基本配置（已完成）

09-11周：完成开发板接口调试。实现视频心率检测，行为模式识别算法构架搭建。

12-14周：训练神经网络训练，测试算法效果，软硬件联合调试。绘制GUI并给出提示信息和统计信息。

15-16周：总结、完成前述中潜在的遗漏部分，尽可能完成上述提到的提高部分。进行报告撰写和PPT制作。

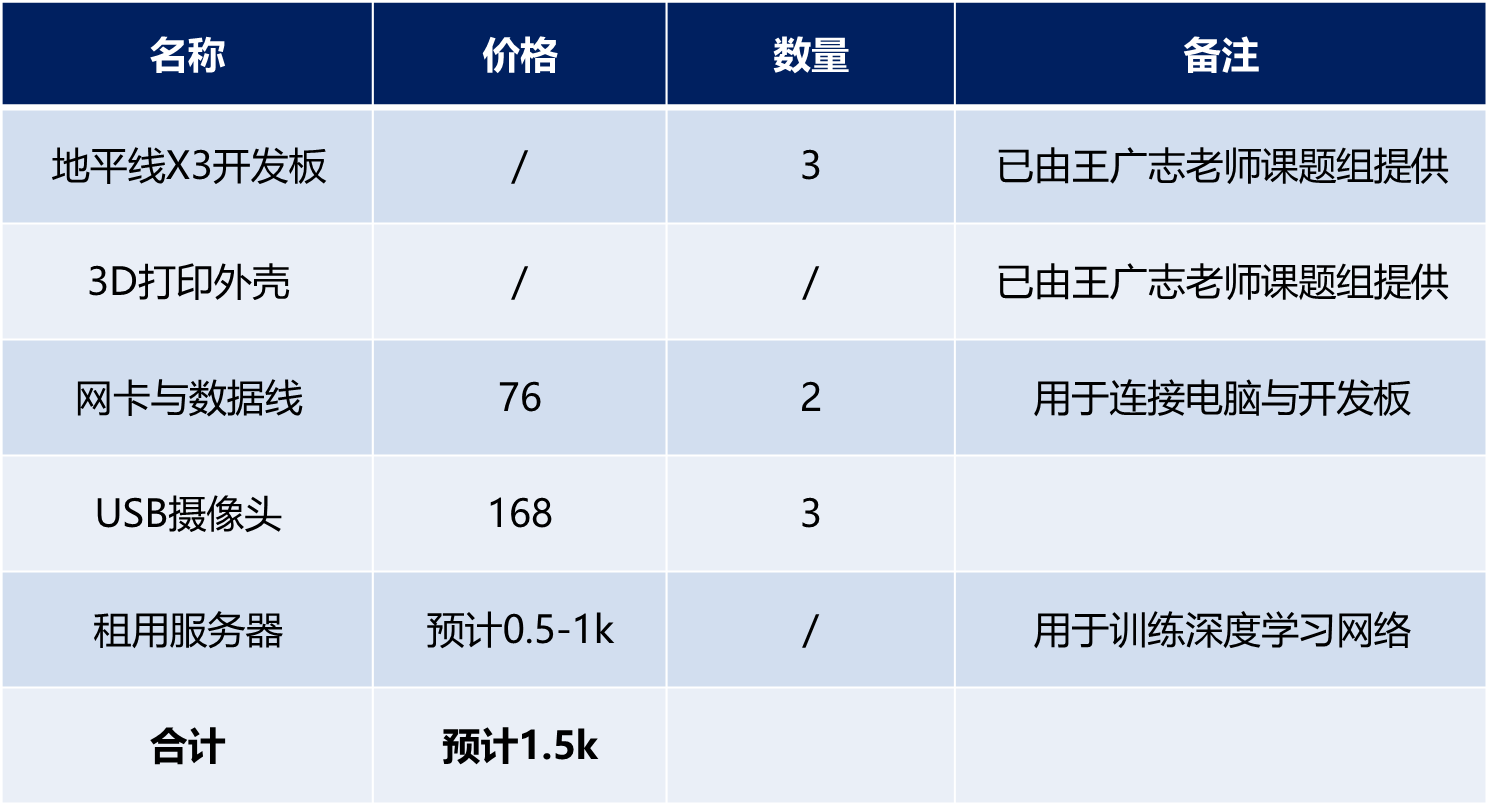
# 七、小组分工计划

李浩伟：完成地平线X3开发板相关硬件调试，实现人体骨架数据提取。

李子涵：调研人体姿态识别相关文献，实现深度学习算法及网络训练。

江柔蓝：完善项目背景和相关调研，实现视频心率检测算法。

# 八、经费预算说明



# 九、参考文献

[1]国家统计局.中华人民共和国2019年国民经济和社会发展统计公报[N]. 2020-2-28.

[2]胡双杰. 基于惯性传感器的实时跌倒检测算法研究与系统设计[D].上海交通大学,2019.

[3]吴聪. 基于惯性传感器的老年人姿态监测系统设计[D].山东大学,2019.

[4]S. Sanyal and K. K. Nundy, "Algorithms for Monitoring Heart Rate and Respiratory Rate From the Video of a User’s Face," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 6, pp. 1-11, 2018, Art no. 2700111, doi: 10.1109/JTEHM.2018.2818687.

[5]R. Favilla, V. C. Zuccalà and G. Coppini, "Heart Rate and Heart Rate Variability From Single-Channel Video and ICA Integration of Multiple Signals," in IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 6, pp. 2398-2408, Nov. 2019, doi: 10.1109/JBHI.2018.2880097.

[6] Shi L , Zhang Y , Cheng J , et al. Skeleton-Based Action Recognition With Directed Graph Neural Networks[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[7] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, Gang Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis", IEEE Conference on Computer Vision and Pattern Recognition, 2016

[8] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, Alex C. Kot, "NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019. [PDF] [bibtex].