Through-Wall Human Pose Estimation Using Radio Signals

基础知识

corss entropy loss

交叉熵描述了两个概率分布之间的距离，当交叉熵越小说明二者之间越接近。

分类问题常用交叉熵。

分类问题最后必须是 one hot 形式算出各 label 的概率， 然后通过 argmax 选出最终的分类。

如果用 MSE 计算 loss， 通过 Softmax后 输出的曲线是波动的，有很多局部的极值点，即非凸优化问题(non-convex)，虽然MSE是凸函数，但经过Softmax，再经过MSE形成的为非凸函数。

而用 Cross Entropy Loss 计算 loss，就还是一个凸优化问题，用梯度下降求解时，凸优化问题有很好的收敛特性。

时空卷积

使用了3D卷积

问题定义

generating 2D skeletal representations of the joints on the arms and legs, and keypoints on the torso and head

RF信号的特性

1. RF信号的空间分辨率比较低，在深度上有10cm左右，在水平和垂直角度上为15度。
2. 信号穿墙后在人体的反射可以看做是镜面反射，因为该雷达的波长为5cm，人体相对较宽，是一种镜面反射。
3. 与基于camera不同，无线信号有不同的表示（复数）和不同的观察角度（水平和垂直）。

论文中方法

跨模态监督

由于使用RF信号，我们无法获取到数据的真实标签，解决思路是使用目前已有的基于视觉的方法与预测人体姿势。

设计了一个跨模态的学生-教师网络，将同步的视频流和RF信号送入网络，来最小化学生网络与教师网络之间的交叉熵损失。

$$min_S \sum_{(I,R)} L(T(I), S(R))$$

I是教师网络T的输入，R是学生网络S的输入，输入包含水平信号和垂直信号，L为损失函数，损失函数定义为图中每个像素的二元交叉熵损失之和。

从RF信号中提取关键点

问题1: 由于人体对信号的反射是一个镜面反射，因此我们无法从一帧图像就获取到人体的姿势。

原因；在一帧图像中可能会丢失一部分信息，而且RF信号的空间分辨率较低。

解决思路：不使用单个帧作为输入，而是让网络查看帧序列。每一输入一个帧序列来输出关键点，虽然网络是以一堆帧作为输入，但是仍然对每一帧的输入都有一个输出。

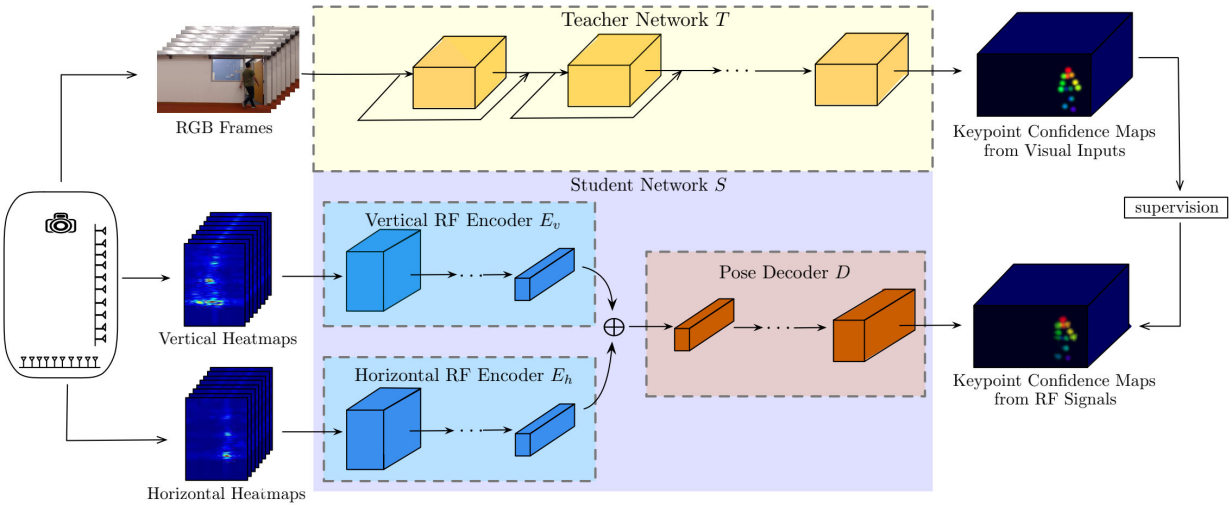
问题2:希望网络在时空变换上是不变的，这样就可以将其从可视的场景推广到穿墙场景

解决思路：使用时空卷积作为学生网络的基本模块。

问题3:在学生网络中需要将RF信号图转换为教师网络的摄像机视图

解决思路：模型必须学习到在原始空间中没有编码的RF信号中的信息表示，然后解码为摄像机视图中的关键点。因此。在学生网络中，（1）对水平和垂直信号分别设置编码器（2）使用一个解码器来得到关键点置信图，输入为水平和垂直信号的编码 。

模型实现和训练



编码网络

每个编码网络的输入是100帧的图像，大致为3.3s（推测其采样率是100Hz）。RF编码网络使用10层9x5x5的时空卷积，步长为1x2x2，激活函数为ReLU。

解码网络

将时空卷积和微条纹卷积结合起来解码姿态，解码器网络有4层，大小为3x6x6，步长为$1 \times \frac{1}{2} \times \frac{1}{2}$，最后一层为$1 \times \frac{1}{4} \times \frac{1}{4}$。激活函数为参数ReLU，最后一层为sigmoid。

训练细节

将实数和虚数在两个channel中输入，batch size为24。

关键点连接

首先对关键点置信度图进行非最大抑制，得到候选关键点的离散峰。为了将不同人的关键点关联起来，使用参考文献[10]（一篇cv的文章）中提到的松弛法，并使用欧氏距离表示两个候选点的权重，根据学习到的关键点置信度图逐帧执行关联，将关键点映射到骨架。

实验

训练时使用70%的可视数据，测试时使用另外30%的可视数据和穿墙数据进行测试。（没有使用穿墙数据进行测试，但是在设计网络时使用时空卷积进行了推广）

预测头，脖子，肩膀，肘部，手腕，臀部，膝盖和脚踝共14个关键点。

评估参数

使用不同关键点的相似度的AP

Baseline

对于可见和部分遮挡的场景，将RF-Pose与OpenPose进行比较，OpenPose是一种最先进的基于视觉的模型，也充当教师网络。

GroundTruth

对于可视场景，使用与RF传感器相结合的摄像机捕捉的图像手动标注人体姿态。对于穿墙场景，使用文章中描述的8个摄像机系统来提供ground truth。对所有8台摄像机拍摄的图像进行标注，以构建三维人体姿态并将它们投射到与RF同步的摄像机视图上。共标注2000张RGB图像：从可见场景测试集和穿透墙场景数据集中分别随机抽取1000张，并使用它们来测试视觉系统和RF-Pose。

我的问题

- 在不穿墙的情况下，我认为我们的雷达拿到的垂直视图以及水平视图比这篇论文中的要稍微好点，应该也可以做出来。