编号 _____

南京航空航天大学

本科毕业设计（论文）

题　目　基于毫米波雷达的人体动作识别系统

学生姓名	董世晨
学　　号	161810129
学　　院	计算机科学与技术学院/人工智能学院/软件学院
专　　业	计算机科学与技术专业
班　　级	1618001
指导教师	张玉书教授

二〇二二年六月

南京航空航天大学

本科毕业设计（论文）诚信承诺书

本人郑重声明：所呈交的毕业设计（论文）是本人在导师的指导下独立进行研究所取得的成果。尽我所知，除了文中特别加以标注和致谢的内容外，本设计（论文）不包含任何其他个人或集体已经发表或撰写的成果作品。对本设计（论文）所涉及的研究工作作出贡献的其他个人和集体，均已在文中以明确方式标明。

作者签名：

日　　期：　20　　年　　月　　日

南京航空航天大学

毕业设计（论文）使用授权书

本人完全了解南京航空航天大学有关收集、保留和使用本人所送交的毕业设计（论文）的规定，即：本科生在校攻读学位期间毕业设计（论文）工作的知识产权单位属南京航空航天大学。学校有权保留并向国家有关部门或机构送交毕业设计（论文）的复印件和电子版，允许论文被查阅和借阅，可以公布论文的全部或部分内容，可以采用影印、缩印或扫描等复制手段保存、汇编论文。保密的论文在解密后适用本声明。

论文涉密情况：

☐ 不保密

☐ 保密，保密期（起讫日期：　　）

作者签名：　　导师签名：

日　　期：　20　　年　　月　　日　　日　　期：　20　　年　　月　　日

摘　　要

近年来，随着机器学习技术的发展，人体动作识别逐渐成为人机交互的一个关键部分，在自动驾驶、智能家居等领域有着广泛应用。毫米波雷达可以有效地解决传统摄像头面临的隐私问题、光照问题，因此在人体动作识别领域受到越来越多的关注。

本文根据 FMCW 雷达的工作原理，通过距离 FFT、计算距离-方位角热力图、多普勒 FFT 等步骤，实现了利用雷达原始数据生成三维点云，生成的每个点包含距离、角度、信号强度、速度等五个维度的信息。

本文为了识别人体动作的三维点云，提出了一套由点云特征提取层、时间序列处理层和全连接分类层组成的机器学习模型架构，每个部分可以独立变换组合。在点云特征提取层中使用基于体素的三维卷积和 PCA 方式、PointNet 和 PointNet++来提取空间域特征；在时间序列处理层中使用 RNN、GRU、LSTM 来提取时间域特征。

本文比较了不同组合的模型在不同数据集上的准确率、模型大小和推理速度，在 RadHAR 和 Pantomime 数据集上分别达到了 97.3%和 95.2%的准确率。本文探究了特征通道、数据增强对模型性能的影响，实验结果表明：使用距离和速度特征与只使用三维坐标相比，准确率各能提升 3%；适当的数据增强也可以提高准确率 4.5%。本文也测试了模型在不同采集环境、角度、速度下泛化能力的表现，实验结果表明：对不同采集角度和不同采集速度的泛化能力比较令人满意，但在不同采集环境下的泛化能力有待提高。

关键词：人体动作识别，毫米波雷达，深度学习，三维点云

ABSTRACT

Recently, with the development of machine learning technology, human gesture recognition has gradually become a key part of human-computer interaction and has been widely used in autonomous driving, smart home, etc. Being able to effectively solve the privacy problems and lighting problems faced by visual cameras, millimeter-wave radar has received more and more attention in this field.

This paper, according to the working principle of FMCW radar, through the steps of range FFT, calculation of distance-azimuth heat map, Doppler FFT, etc., implements an algorithm to generate 3D point clouds from the raw data collected by radar. Each generated point contains information on distance, angle, signal strength, speed, etc.

To identify the 3D point cloud of human gestures, this paper proposes a machine learning model consisting of a point cloud feature extraction layer, a time series processing layer, and a fully connected classification layer. The models used in each layer can be independently chosen and combined. In the point cloud feature extraction layer, voxel-based 3D convolution and PCA methods, PointNet, and PointNet++ are used to extract features in the spatial domain; in the time series processing layer, RNN, GRU, LSTM are used to extract features in the time domain.

This paper compares the accuracy, model size, and inference speed of different combinations of models on different datasets, it achieves 97.3% and 95.2% accuracy on RadHAR and Pantomime datasets, respectively. This paper explores the influence of feature channels and data augmentation on model performance and concludes that: distance and speed features can each improve the accuracy by 3% compared to models using only xyz features. This paper also tests the model's generalization ability under different acquisition environments, angles, and speeds, and concludes that: distance and speed features can each improve the accuracy by 3%; appropriate data augmentation can improve the accuracy by 4.5%; the generalization ability for different acquisition angles and different acquisition speeds is satisfactory, but that in different acquisition environments remains to be improved.

KEY WORDS: Human gesture recognition, Millimeter wave radar, Deep learning, 3D point cloud

目录

第一章 绪论

1.1 背景和意义

人体动作识别一直是人机交互的一个重要组成部分[1]，在自动驾驶汽车的交警手势识别系统[2]、独居老人/残疾人的医疗监护系统[3]、健身房人员的安全保障系统[4]、工厂或超市的人员监视系统[5]等领域有着广泛的应用。

传统的人体动作识别系统一般是基于视觉摄像头[6]或可穿戴设备[7]实现的。视觉摄像头虽然有着最高的识别精度，但面临着一定程度的隐私问题[8]，并且必须在满足光线充足、视野清晰没有遮挡、天气晴朗无雾等条件的环境下工作；可穿戴设备虽然可以去除对光照情况的要求，但其严苛的设备要求使得在实际使用中受到许多限制[9]而不那么方便。

相比之下，毫米波雷达既可以避免隐私问题，也可以在完全无光的室内，能见度很低的大雾和雨天，甚至是被遮挡的情况下工作[10]，同时还不需要被试者穿戴额外的设备。得益于此，基于毫米波雷达的人体动作识别系统在近年来得到越来越多的关注和研究。

1.2 国内外研究现状

1.2.1 人体动作信号采集技术

为了识别人体动作，除了穿戴式设备，研究者们还尝试了使用 RGB 视觉摄像头、深度摄像头，以及利用 WiFi/雷达频率信号进行无线感知等方式。基于机器视觉的方式已经可以做到利用视频生成高精度人体骨架和三维点云[11]，而在无线感知方向：Jaime Lien 等人提出了 Soli[12,13]，一种基于毫米波雷达的高精度的微型手势识别系统；Tianhong Li 等人将雷达和视觉摄像头相结合，提出了一个神经网络模型，实现了将雷达频率信号转换为人体骨架信息[14,15]；Yongsen Ma 等人使用 WiFi 的信道状态信息，提出了一个基于卷积神经网络的手语识别系统 SignFi[16]；Sameera Palipana 等人提出了一套基于毫米波雷达的，使用 PointNet++和 LSTM 的空中手势识别系统 Pantomime[17]。

在硬件层面，美国德州仪器（Texas Instruments）公司在 2018 年推出的 IWR1642 型号的商用毫米波雷达体积小、集成度高、使用方便，迅速推动了在自动驾驶、智能家居等领域使用毫米波雷达进行人体动作识别的发展[18]。

1.2.2 人体动作三维点云识别技术

LiuHao Ge 等人[19]将三维点云投影到多个平面上，在每个平面上应用卷积神经网络。Huang Su 等人[20]也提出了一种多视角的卷积神经网络来识别三维模型。J. Owoyemi 等人

[21]、A. D. Singh 等人[22]、Peijun Zhao 等人[23]在各自提出的方法中都使用了将三维点云体素化后使用三维卷机神经网络来进行人体手势识别的方式。但是，投影和体素化这两种方法都面临着将点云数据预处理带来的信息丢失，其模型效果也不尽人意。

Charles R. Qi 和 Hao Su 等人在 2017 年创新性地提出了 PointNet[24]和 PointNet++[25]两个模型架构，该模型可以直接使用三维点云作为模型的原始数据，相比于之前常用的三维点云体素化的方式，避免了体素化带来的信息损失，大幅提高了模型处理精密的三维点云的能力，使得在三维点云上训练出能够识别复杂人体动作的模型成为可能。

1.3 本文主要工作

本文的工作可以概括为以下三点：

（1） 本文实现了使用 FMCW 雷达给出的原始数据生成三维点云，点云中每个点包含距离、方位角、俯仰角、雷达信号强度、径向速度这五个用于描述物体在三维空间信息的数据；

（2） 本文提出了一套由点云特征提取层、时间序列处理层和全连接分类层组成的机器学习模型架构，来对人体动作的三维点云时间序列进行识别和分类；

（3） 本文比较了基于体素的三维卷积和 PCA 方法、PointNet、PointNet++（SSG）和 PointNet++（MSG）这五种方法在 RadHAR 和 Pantomime 数据集上的准确率、模型大小和推理速度，并探究了特征通道、数据增强对模型性能的影响，以及模型在不同背景环境、不同采集角度、不同动作执行速度下的表现。

1.4 论文组织结构

本论文的组织结构如下：

第一章 介绍研究背景、意义和研究现状，阐述本文的主要工作；

第二章 介绍毫米波雷达的工作原理和三维点云的生成算法流程；

第三章 阐述用于识别人体动作三维点云时间序列的机器学习模型架构；

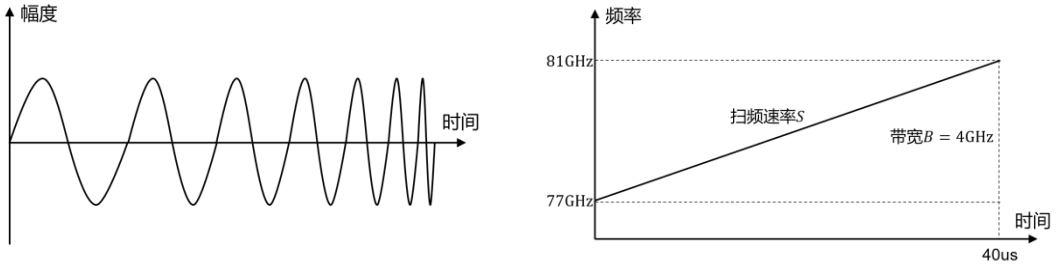
第四章 对不同模型组合的性能进行实验，分析不同因素对准确率的影响；

第五章 总结本文的研究成果，对未来的研究做出展望。

第二章 毫米波雷达的工作原理及三维点云的生成

2.1 FMCW 雷达的工作原理

调频连续波（Frequency-Modulated Continuous-Wave，FMCW）雷达可以通过快速发射一组调频信号（chirp），并比较发射信号与由待测物体反射回来的反射信号之间的差异，从而推算出多个待测物体相对于雷达的距离、径向速度、方位角等信息[26]。雷达发射出的每一个调频信号都是一个频率不断线性增加的正弦波，如图 2.1（a）所示。我们通常使用频率-时间图描述调频信号，图 2.1（b）展示了带宽为 4GHz 的 77GHz 毫米波雷达发出的调频信号的频率-时间图。



（a） （b）

图 2.1 调频信号的幅度-时间图和频率-时间图

调频信号通过发射天线发射出去后，经待测物体反射后，被雷达的接收天线阵列接收。反射信号将同发射信号一起，送入混频器中。FMCW 雷达中的混频器输出的信号的频率和相位均是两个输入信号的频率和相位之差。形式化而言，记发射信号为公式 2.1：

$$x(t) = \exp\left(-i \cdot 2\pi\left(f_0 + \frac{B}{2T}t\right)t\right) \tag{2.1}$$

式中，f_0为初始频率，B为带宽，T为调频信号的持续时间。记反射信号为公式 2.2：

$$y(t) = \sum_{i=1}^{M} A_i \exp\left(-i \cdot 2\pi\left(f_0(t - t_i) + \frac{B}{2T}(t - t_i)^2\right)\right) \tag{2.2}$$

式中，A_i为第i个待测物体的衰减，$t_i = \frac{d_i(t)}{c}$为信号传播时间。则混频器的输出为两者相除，即公式 2.3：

$$\hat{y}(t) = \sum_{i=1}^{M} A_i \exp\left(-i \cdot 2\pi\left(\frac{B}{2T}t_i t + f_0 t_i - \frac{B}{2T}{t_i}^2\right)\right) \tag{2.3}$$

图 2.2（a）使用频率-时间图直观地展示了发射信号$x(t)$和多个待测物体的反射信号$y(t)$，图 2.2（b）他们经过混频器之后的输出信号$\hat{y}(t)$。混频器的输出信号的频率和待测物体的距离成正比关系，因此可以通过对输出信号做快速傅立叶变换（Fast Fourier Transform，FFT），得到的结果中的每一个峰值都对应着一个待测物体。

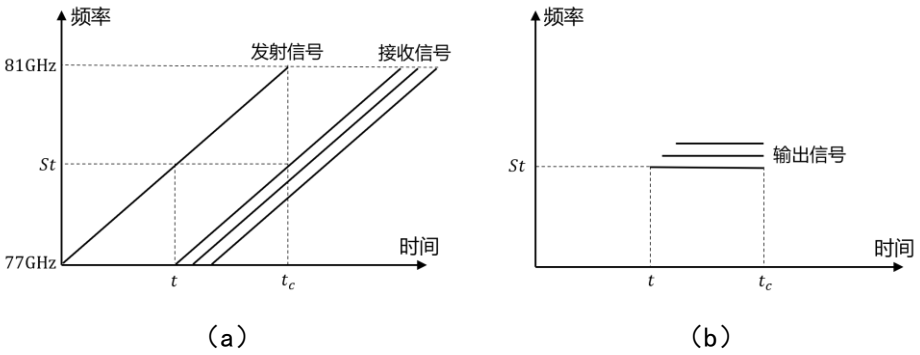


图 2.2 发射信号、接收信号、输出信号的频率-时间图

FMCW 雷达会在数十毫秒内连续发射数百个调频信号，这些调频信号组成一个帧（frame）。由于帧的持续时间较短，待测物体在一帧之内的位移通常不会超过信号的波长，因此可以通过观察一帧之内的不同调频信号中，同一个待测物体的反射信号的相位变化来计算出该待测物体的速度，而不用担心相位差超过2π而发生相位缠绕的现象。

FMCW 雷达通常包含由若干个天线组成的接收天线阵列。当信号到达方向（Angle of Arrival，AoA）与接收天线阵列平面不垂直时，同一个信号到达阵列中不同接收天线时的相位会出现不同程度的延迟，如图 2.3 所示。CAPON 波束形成算法[27]给出了在最小方差无畸变的前提下对信号到达方向的最优估计。

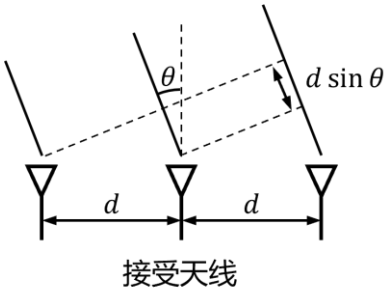


图 2.3 接收信号延迟示意图

本章的剩余篇幅将详细阐述如何本文是如何利用混频器输出信号的这些性质来生成三维点云、以及对点云进行去噪的。

2.2 生成三维点云

本文实现的三维点云生成算法流程如图 2.4 所示。

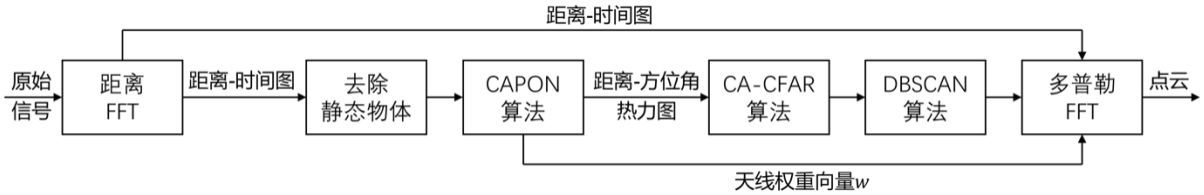


图 2.4 三维点云生成算法流程图

2.2.1 使用的数据集和雷达参数

本结将采用 E. Gambi 等人[28]提供的开源毫米波雷达原始信号数据集来可视化地呈现算法的中间结果。该数据集包含了慢速行走、快速行走、插口袋行走、藏着一个瓶子行走、瘸着行走、摆臂行走这六种步态的毫米波雷达原始波形数据，共 231 条数据，总计达 100.8GB 数据量。表 2.1 描述了采集该数据集时使用的雷达参数设置。

表 2.1 雷达参数

参数	值
接受天线数量	4
每条数据时长	16s
每条数据帧数	400
每帧包含的调频信号数量	128
每个调频信号中采样次数	512
采样频率	10Msps
扫频速度	60.012MHz/μs
使用的带宽	3.6GHz

2.2.2 距离 FFT

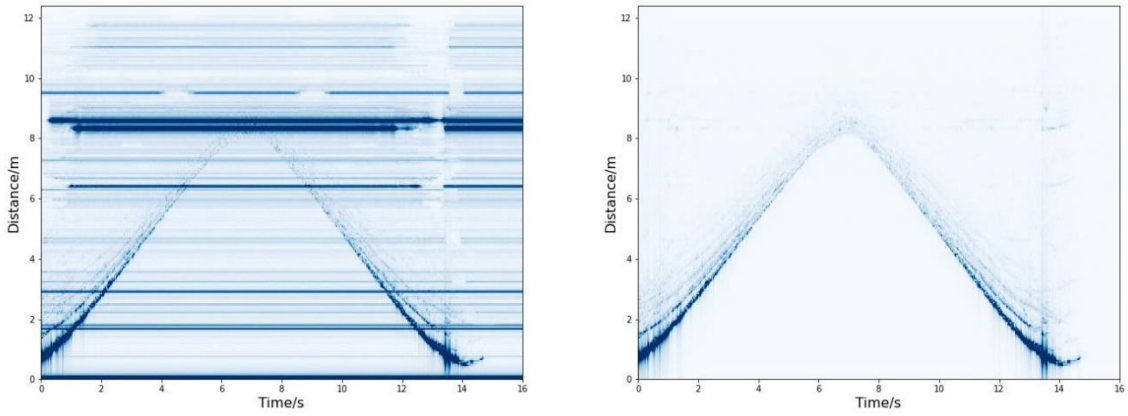
雷达的原始数据是 4×26214400 的复数矩阵，为了便于处理，先将其变形到 4×400×128×512 的四维张量，其中四个维度依次是天线维度、帧维度、调频信号维度以及采样维度。

由于混频器的输出信号中的每个频率与待测物体的距离满足公式 2.4：

$$d = \frac{fc}{2S} \tag{2.4}$$

式中，f为接收信号频率，c为光速，S为扫频速率。

因此本文首先对原始数据在采样维度进行一维 FFT 变换，得到的结果如图 2.5（a）所示，横坐标为时间，纵坐标为待测物体到雷达的距离，该图展示了一个人在 16 秒内先步行远离雷达至距离 8 米处，然后返回至雷达的过程。



（a）去除静态物体前　　（b）去除静态物体后

图 2.5 距离-时间图

2.2.3 去除静态物体

图 2.5（a）所展示的结果中存在大量的静态物体，通常为墙壁、地面等，它们不是观测的主要对象，但会非常干扰后续处理，因此需要去除。

由于静止物体在经过多谱勒 FFT 后都落于速度为 0 的原点上，将原点的幅度改为 0，即可去除静止物体。数学上可以证明，FFT 后的结果的原点等于数据的平均值，因此只需将距离 FFT 之后的结果减去它在调频信号维度的平均值，就可以做到去除静态干扰物体的效果。图 2.5（b）展示了该数据去除静态物体后的结果。

2.2.4 使用 CAPON 波束形成算法计算距离-方位角热力图

如图 2.3 所示，对于某个特定信号到达角度θ，可以对阵列中的每个接收天线i，设置信号延迟a_i，从而抵消掉信号到达角度与阵列平面不垂直带来的信号延迟问题，使得来自于该方向的信号互相重叠加强，来自于其他方向的信号彼此抵消。数学上可以将每个天线接收到的信号乘以导引向量$\boldsymbol{a}$（steering vector）来改变每个天线的信号的相位。对于信号到达角度θ，导引向量$\boldsymbol{a}$的构造如公式 2.5 所示：

$$\boldsymbol{a}_\theta = \begin{bmatrix} 1 & e^{2i\pi\frac{d}{\lambda}\sin\theta} & \cdots & e^{2(N_r-1)i\pi\frac{d}{\lambda}\sin\theta} \end{bmatrix}^T \tag{2.5}$$

式中，d为接受天线间距，λ为信号波长，N_r为接受天线数量。

为了得到导引向量$\boldsymbol{a}$下的信号强度P和各天线权重$\boldsymbol{w}$，Capon 在 1969 年提出了著名的最小方差无畸变波束形成算法[27]（Minimum Variance Distortionless Response，MVDR），该算法计算得到的各天线权重$\boldsymbol{w}$能够使得在信号增益恒为 1 的情况下，随机噪声的方差最小，公式 2.6 是该算法的形式化表示：

$$\min_{\boldsymbol{w}} \boldsymbol{w}^H \mathbf{R} \boldsymbol{w} \text{ 满足 } \boldsymbol{w}^H \boldsymbol{a} = \mathbf{1} \tag{2.6}$$

式中，$\boldsymbol{R}$为噪声的协方差矩阵，使用接收到的雷达信号$\boldsymbol{X}$来估计，如公式 2.7 所示：

$$\boldsymbol{R} = \frac{1}{N_c} \sum_{c=1}^{N_c} \boldsymbol{X}_c \boldsymbol{X}_c^H \tag{2.7}$$

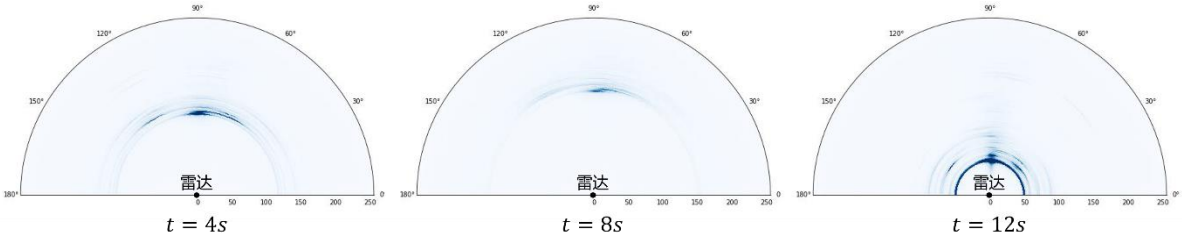
式中，N_c为每帧内调频信号的数量，在实际应用中，为了保证数值稳定性，常对它加上一个单位矩阵的α倍。

经过一系列数学推导，各天线权重$\boldsymbol{w}$和信号强度P的计算公式如公式 2.8 和公式 2.9 所示：

$$\boldsymbol{w}_\theta = \frac{\boldsymbol{R}^{-1} \boldsymbol{a}_\theta}{\boldsymbol{a}_\theta^{\mathrm{H}} \boldsymbol{R}^{-1} \boldsymbol{a}_\theta} \tag{2.8}$$

$$P_\theta = \frac{1}{\boldsymbol{a}_\theta^{\mathrm{H}} \boldsymbol{R}^{-1} \boldsymbol{a}_\theta} \tag{2.9}$$

本文设定角度分辨率为 180，对于雷达每一帧的不同距离的信号，构造出-90°到 90°之间共 180 个角度的导引向量，对它们分别应用一遍 CAPON 波束形成算法，计算出它们的信号强度P，得到每一帧的距离-方位角热力图，如图 2.6 所示。该步骤计算量较大，占了三维点云生成的算法流程中大部分时间。



$t = 4s$　　$t = 8s$　　$t = 12s$

图 2.6 CAPON 算法的结果

2.2.5 使用 CA-CFAR 算法对热力图去噪

上一步生成的距离-方位角热力图存在大量噪声，并不能很好地显示出人物的具体位置信息，因此本文采用了恒虚警率检测算法[29]（Cell-Average Constant False Alarm Rate，CA-CFAR）来找出真正属于待测物体的信号点。

该算法认为如果一个点的值显著地大于它附近的若干个点的平均值，则它大概率是真实的点。为了避免信号范围覆盖了多个点，该算法计算附近点的平均值时不包含紧挨着中心点的若干个点，图 2.7 形象地说明了该算法的原理。

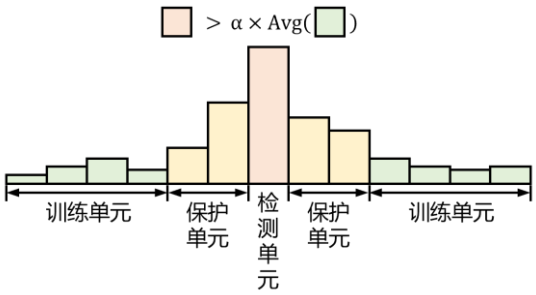


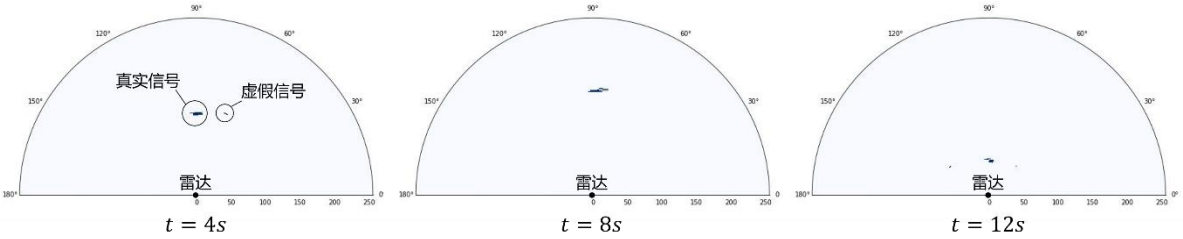
图 2.7 CA-CFAR 算法示意图

该算法通过公式 2.10 来设置被认为是真实点的值高于附近点的平均值的倍率阈值α，保证了虚警率（即算法输出的真实点实际不存在的概率）为恒定的，在本文中设定为10^{-3}。

$$\alpha = N\left(P_f{}^{-\frac{1}{N}} - 1\right) \tag{2.10}$$

其中P_f为虚警率，N为用于计算附近平均值的点的数量。

本文采用两遍一维的 CFAR 算法来实现对二维的热力图去噪，即沿距离维度和角度维度分别进行一遍一维的 CFAR 算法，只有在两次 CFAR 算法中都被认定为是真实的点才会被保留下来。对于距离和角度两个维度，本文都采用了双边各 24 个点用于计算平均值，不包含紧挨着中心点的双边各 8 个点，得到的结果如图 2.8 所示。



$t = 4s$　　$t = 8s$　　$t = 12s$

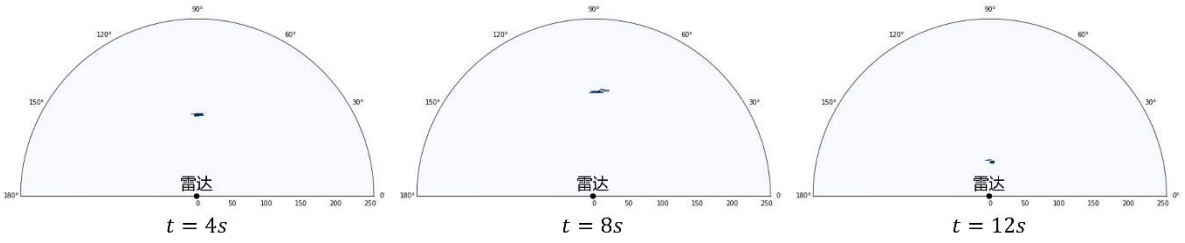
图 2.8 CA-CFAR 算法的结果

2.2.6 使用 DBSCAN 算法去除离群点

虽然 CFAR 算法给出的结果的虚警率极低，但其中仍然存在一些噪音，这可能是因为雷达信号经墙壁等平面反射得到的。本文采用 DBSCAN 聚类算法[30]（Density-Based Spatial Clustering of Applications with Noise）来去除 CFAR 算法给出的点集中的离群点。

该算法是一种基于密度的聚类算法，在聚类的同时还能鉴别出数据中的离群点。首先找出所有在ε半径内有超过m个点的核心点，彼此距离不超过ε的核心点构成了若干个连通图，每个连通图中所有核心点附近ε半径内的所有点构成了一个类别，那些没有被划分进任何类别的点为离群点。

本文只保留了该算法输出类别中点数量最多的那个类别，得到的结果如图 2.9 所示。



$t = 4s$　　$t = 8s$　　$t = 12s$

图 2.9 DBSCAN 算法的结果

2.2.7 使用多普勒 FFT 计算径向速度

除了获取每个点的坐标，FMCW 雷达还可以通过观察同一帧之内的不同调频信号中，每个点对应的反射信号的相位变化来计算出该点的径向速度，相位变化的角速度ω与该点的径向速度v之间的关系如公式 2.11 所示：

$$v = \frac{\lambda\omega}{4\pi T_c} \tag{2.11}$$

式中，λ为信号波长，T_c为一个调频信号的时长。

为了精确地获取到来自于某个点的帧信号，本文采取以下两种方法：1）利用距离 FFT 给出的结果，选取其中该点对应距离处的信号；2）利用 CAPON 算法在该点的方位角处计算出的各天线权重$\boldsymbol{w}$（即公式 2.5），选取该点对应方位角处的信号。结合以上两种方式，就可以获取到给定距离和方位角下的雷达信号，对该信号按调频信号维度进行一维 FFT 变换，即可得到该点处的速度频谱图，本文取幅度最大处的速度作为该点的径向速度。

由于该步骤的计算量也比较大，本文仅对 DBSCAN 算法给出的真实点做多普勒 FFT，得到的效果如图 2.10 所示。可以看到，当人物远离雷达行走时，速度为负（显示为红色），靠近雷达行走时，速度为正（显示为蓝色）。

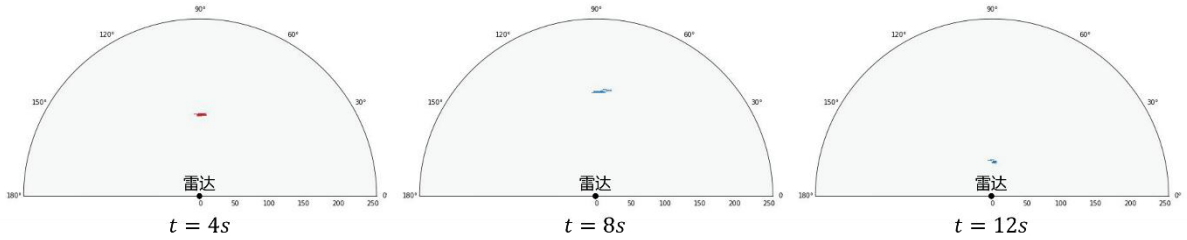


图 2.10 多普勒 FFT 后的结果

2.3 本章小结

根据 FMCW 毫米波雷达的工作原理，本文通过距离 FFT、去除静态物体、计算距离-方位角热力图、热力图去噪、去除离群点、多普勒 FFT 这六个步骤，实现了使用 FMCW 雷达给出的原始数据生成三维点云，点云中每个点包含距离、方位角、俯仰角、雷达信号强度、径向速度这五个用于描述物体在三维空间信息的数据。

值得注意的是，本章仅展示了距离和方位角这两个维度的“二维点云”的生成流程，一是因为使用的数据集的天线阵列只有一个维度，只能生成二维的点云；二是为了便于可视化地展示算法的中间结果。在此基础上实现生成三维点云是容易的，只需要在另一个维度上再次执行一遍 CAPON 算法即可。

第三章 基于深度神经网络识别人体动作三维点云

3.1 模型简介

3.1.1 模型的输入和输出

本章提出一套深度神经网络模型来完成对人体动作三维点云的分类任务。该模型的输入是一组三维点云的时间序列，由被试者在雷达前做出不同动作（如交警的各种指挥手势）采集得到；该模型的输出是该动作的类别。

3.1.2 任务难点及解决方案

相比于其他机器学习任务，对于基于 FMCW 雷达采集的人体动作三维点云的分类任务面临以下三个难点：

（1） 三维点云是一个点集的无序序列，任意交换其中的元素顺序不应该对分类结果有影响，即神经网络不应该对点集元素的顺序敏感；

（2） 每一条训练数据由多个三维点云组成时间序列，该时间序列的顺序会对分类结果产生影响（如顺时针摆动手臂与逆时针摆动手臂的动作应属于不同类别），即神经网络应该对时间序列的顺序敏感；

（3） 由于 FMCW 雷达的特性，点云通常是非常稀疏的，并且点的分布密度是不均匀的，距离雷达越远，点分布地越稀疏。

对于第一个难点，本文采取了点云转换为体素的方法以及 PointNet[24]/PointNet++[25]这两种方式来解决；对于第二个难点，本文采用了 GRU[31]、LSTM[32]等专用于处理序列数据的网络结构来解决；对于第三个难点，本文通过合并帧、合理选取体素大小、PointNet++的多尺度分组等方法解决。

3.1.3 模型架构

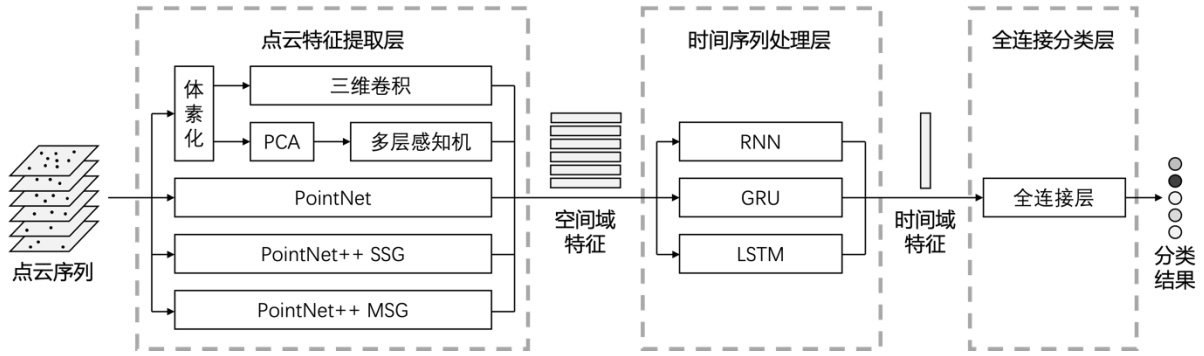


图 3.1 本文模型架构图

如图 3.1 所示，本文将模型分为了点云特征提取层、时间序列处理层、全连接分类层三个相互解耦的模块，每个模块间的接口（即输入输出）是固定的，它们的具体实现可以独立变化。

点云特征提取层是本章的主要部分，用于提取单个点云的空间域特征，输入是一个的无序点集，每个点包含若干维的特征值，输出是一个可以描述该点云空间域特征的向量。该层的实现包含基于体素的三维卷积模型、对体素使用主成分分析后进行全连接的模型、PointNet、PointNet++（单尺度分组）、PointNet++（多尺度分组）这五种模型。

时间序列处理层用于提取多个点云的时间域特征，输入是一组时间序列中每个点云经过点云特征提取层得到的特征向量序列，输出是一个可以描述该点云时间序列的空间域特征的向量。该层的实现包含 RNN、GRU、LSTM 这三种模型。

全连接分类层是一个简单的多层感知机，输入是时间序列处理层给出的空间域特征向量，输出是各个类别的置信度。

本章的剩余部分将依次介绍这些模型。

3.2 基于体素的两种模型

3.2.1 三维点云体素化

在 PointNet 和 PointNet++被提出前，将三维点云体素化是比较常用的方式。体素化是指将三维空间划分为很多个立方体单位（称为体素），统计落入每个体素的点的数量，即可得到该点云在三维空间不同位置的密度。对于那些每个点有附加特征值（如速度、雷达信号强度等）的点云，每个体素的特征值由落入该体素的点的特征值取平均得到。点云体素化的流程与点云中点的顺序无关，因此自然地解决了 3.1.2 节提到的难点 1。

每个体素的大小、体素化的边界等参数对于模型效果会产生关键的影响。如果单个体素太大，就会丢失掉每个体素内部点的精密结构信息，导致模型效果下降；如果单个体素太小，数据量将呈立方数量级倍增，大大增加计算量，且每个体素只能包含极少的点甚至不包含任何点，不利于模型提取局部特征。

3.2.2 三维卷积模型

由于体素数据格式和二维 RGB 图像十分类似，有良好的平移不变性，且同样拥有多个通道（如密度通道、速度通道等），本文参考广泛应用于图像的二维卷积层，尝试使用了三维卷积层来对体素进行特征提取。本文设计的三维卷积模型包含若干个依次由三维卷积层、Batch Normalization 层、Max Pooling 层组合而来的卷积模块。

3.2.3 使用 PCA 对体素进行降维

由于体素数据存在大量冗余，例如大部分数据集中在中心部分，靠近边缘的体素通常不包含任何点，A. Singh 和 S. Sandha 等人[22]在 2019 年提出可以使用主成分分析法（Principle Component Analysis，PCA）对体素数据进行降维。首先将三维的多通道体素数据展平成一维向量，使用 PCA 对该向量进行降维，随后送入一个多层感知机来提取该体素的特征向量。

3.3 PointNet 和 PointNet++

3.3.1 PointNet

Charles R. Qi 和 Hao Su 等人在 2017 年创新性地提出了 PointNet[24]，该模型直接接受三维点云形式的输入，避免了点云体素化带来的细节丢失，在 ModelNet40 数据集上达到了 89.2%的准确率。

PointNet 是通过对点云集合中的每个点执行相同的操作（如旋转变化、多层感知机等），并通过使用对输入参数的顺序和数量不敏感的对称函数（如求最大值函数），来使得 PointNet 对于三维点云的点集数量和顺序不敏感。

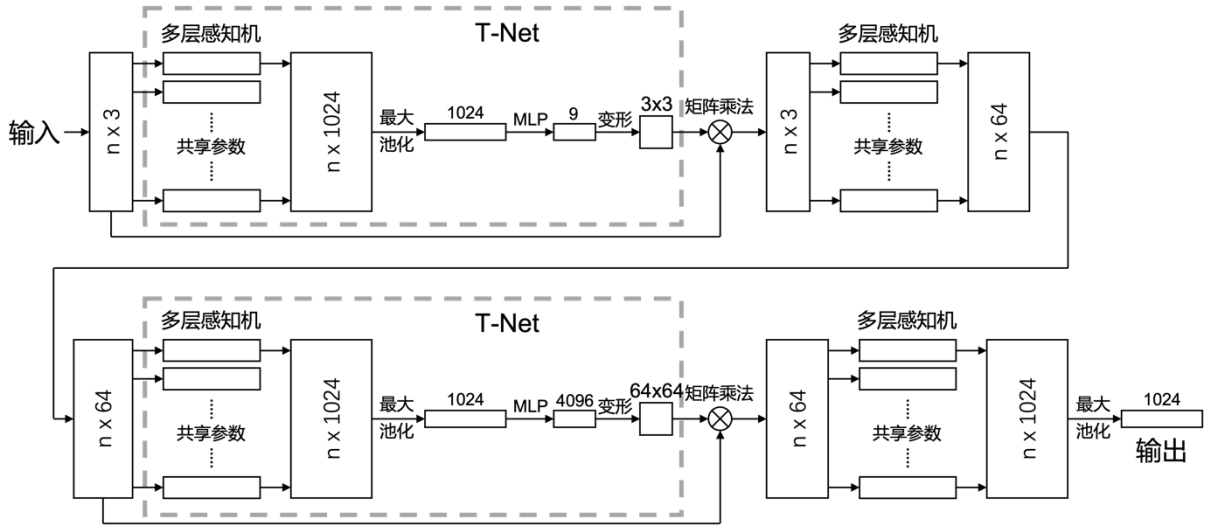


图 3.2 PointNet 模型架构图

图 3.2 展示了 PointNet 的结构，它由若干个重复的模块和一个最大池化层组成。每个模块包含一个 T-Net 和一个多层感知机：

（1） T-Net 用于将输入的点集进行旋转变换，其结构类似于一个只有一个模块的迷你 PointNet，输出一个大小等同于输入通道数的方阵，代表了一个旋转矩阵，将其应用于输入的点集，该文的作者认为可以实现点集在特征空间中对齐。

（2） 多层感知机将应用于输入点集的每一个点，其权重是共享的。实现上可以使用多个核大小为 1 的一维卷积层来达到该效果。

（3） 最后的最大池化层取输出点集中每个通道的最大值，作为输入点集的全局特征向量。

3.3.2 PointNet++单尺度分组

PointNet 虽然首次实现了直接对三维点云数据进行分类，但是它的网络结构使它只能在整个点集上进行推理，无法提取到局部的精细特征。Charles R. Qi 等人因此在同年改进了 PointNet，提出了 PointNet++[25]。该模型使用了一个类似于卷积核的结构来实现对点云局部精细结构进行特征提取，有效提高了对复杂模型的特征提取能力，在 ModelNet40 数据集上达到了 91.9%的准确率。

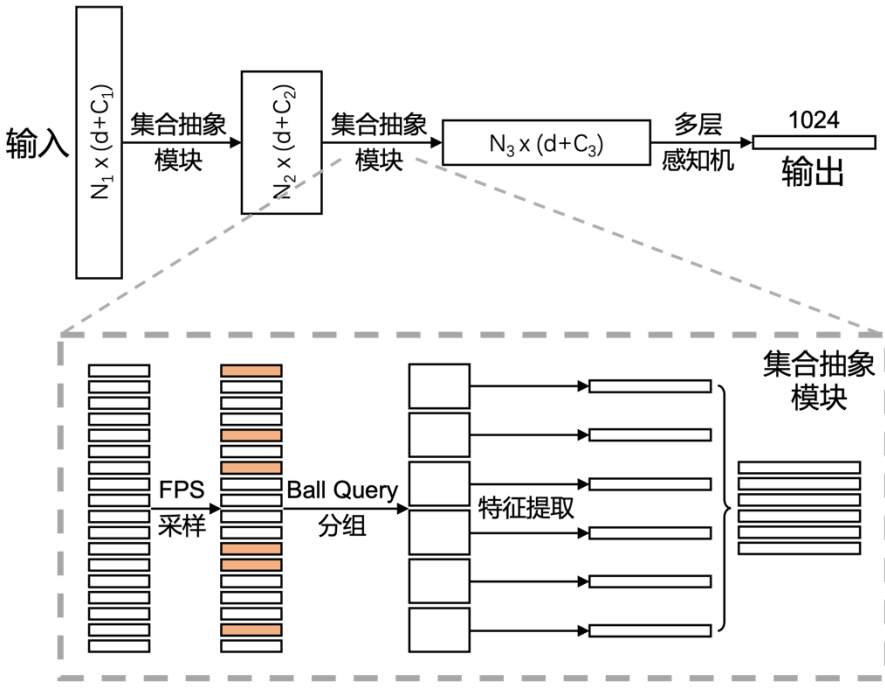


图 3.3 PointNet++模型架构图

图 3.3 展示了 PointNet++的结构，它由多个集合抽象模块（Set Abstraction）组成，每个集合抽象模块包含采样、分组、特征提取三个步骤：

（1） 采样层旨在对输入点集进行密度无关的均匀采样。PointNet++使用了最远点采样算法（Farthest Point Sampling，FPS），该算法每次选取一个距离已选择点集距离最远的点。相比于随机选择，FPS 算法可以很好地覆盖那些点密度很低的区域。

（2） 分组层旨在抽取出采样点附近的局部点集。对于采样层采样得到的每个点，找出它半径r内的最近的k个点（称为 Ball Query 算法），并计算这些点相对于采样点的相对坐标，作为输出。

（3） 特征提取层旨在提取每个局部点集的特征向量。该层延续了 PointNet 的设计，将多层感知机应用于局部点集中的每个点。值得注意的是，PointNet++不再使用 T-Net 对点集进行旋转，并且保留了每个点的原始三维坐标。

宏观地看集合抽象模块，它相当于使用了更少但更高维的点，来描述输入点集中更多但更低维的点。

3.3.3 PointNet++多尺度分组

为了解决 FMCW 雷达产生的三维点云中点的分布密度不均匀的问题，Charles R. Qi 等人提出了多尺度分组（Multi-Scale Grouping，MSG）和多分辨率分组（Multi-Resolution Grouping，MRG）两种方法[25]，本文主要实现了前者。

MSG 方法的核心思想是在分组层选取多个不同尺度的半径r和数量k，得到的局部点集分别送入特征提取层中提取特征，最后将多个尺度的特征连接起来，作为可以描述该采样点附近不同尺度下的特征的向量。MSG 方法和 3.3.2 小节介绍的单尺度分组（Single-Scale Grouping，SSG）方法的对比如图 3.4 所示。

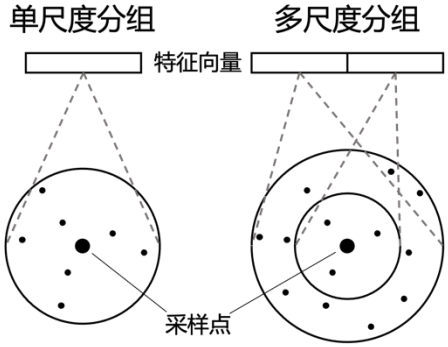


图 3.4 PointNet++的单尺度分组与多尺度分组的对比

3.4 时间序列模型

本文尝试使用了三种常见的用于处理序列数据的神经网络结构来提取人体动作三维点云序列的时间域特征，分别为循环神经网络、门控循环单元、长短期记忆网络。

3.4.1 循环神经网络 RNN

循环神经网络（Recurrent Neural Network，RNN）作为处理序列数据最基础的结构，仅包含一个隐藏状态。每次循环，根据上一时刻的隐藏状态和当前时刻输入的序列元素，计算出当前时刻的隐藏状态，它的隐藏状态和输出的更新公式如公式 3.1 和公式 3.2 所示：

$$\boldsymbol{h}_t = \sigma(\boldsymbol{W}_{\boldsymbol{h}}\boldsymbol{x}_t + \boldsymbol{U}_{\boldsymbol{h}}\boldsymbol{h}_{t-1} + \boldsymbol{b}_{\boldsymbol{h}}) \tag{3.1}$$

$$\boldsymbol{y}_t = \sigma\left(\boldsymbol{W}_{\boldsymbol{y}}\boldsymbol{h}_t + \boldsymbol{b}_{\boldsymbol{y}}\right) \tag{3.2}$$

式中，$\boldsymbol{h}$是隐藏状态，$\boldsymbol{x}$是输入向量，$\boldsymbol{y}$是输出向量，$\boldsymbol{W}$、$\boldsymbol{U}$、$\boldsymbol{b}$是训练参数。

3.4.2 门控循环单元 GRU

门控循环单元[31]（Gated Recurrent Unit，GRU）通过引入更新门和重置门来控制记忆的更新和遗忘，来解决 RNN 无法保持长期记忆以及梯度消失的问题。它的更新门、重置门和隐藏状态的更新公式如公式 3.3～公式 3.5 所示：

$$\boldsymbol{z}_t = \sigma(\boldsymbol{W}_{\boldsymbol{z}}\boldsymbol{x}_t + \boldsymbol{U}_{\boldsymbol{z}}\boldsymbol{h}_{t-1}) \tag{3.3}$$

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_{\boldsymbol{r}}\boldsymbol{x}_t + \boldsymbol{U}_{\boldsymbol{r}}\boldsymbol{h}_{t-1}) \tag{3.4}$$

$$\boldsymbol{h}_t = (1 - \boldsymbol{z}_t) * \boldsymbol{h}_{t-1} + \boldsymbol{z}_t * \tanh\left(\boldsymbol{W}_{\boldsymbol{h}}\boldsymbol{x}_t + \boldsymbol{U}_{\boldsymbol{h}}(\boldsymbol{r}_t * \boldsymbol{h}_{t-1})\right) \tag{3.5}$$

式中，$\boldsymbol{h}$是隐藏状态，$\boldsymbol{x}$是输入向量，$\boldsymbol{z}$是更新门，是$\boldsymbol{r}$遗忘门，$\boldsymbol{W}$、$\boldsymbol{U}$是训练参数，$*$是按元素乘。

3.4.3 长短期记忆网络 LSTM

长短期记忆网络[32]（Long Short-Term Memory，LSTM）使用了输入门来决定输入数据的重要程度、输出门来决定是否使用隐藏状态来输出、遗忘门来决定是否衰减隐藏状态表示的记忆。它的输入门、输出门、遗忘门和隐藏状态的更新公式如公式 3.6～公式 3.10 所示：

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}_{\boldsymbol{f}}\boldsymbol{x}_t + \boldsymbol{U}_{\boldsymbol{f}}\boldsymbol{h}_{t-1} + \boldsymbol{b}_{\boldsymbol{f}}\right) \tag{3.6}$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{\boldsymbol{i}}\boldsymbol{x}_t + \boldsymbol{U}_{\boldsymbol{i}}\boldsymbol{h}_{t-1} + \boldsymbol{b}_{\boldsymbol{i}}) \tag{3.7}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{\boldsymbol{o}}\boldsymbol{x}_t + \boldsymbol{U}_{\boldsymbol{o}}\boldsymbol{h}_{t-1} + \boldsymbol{b}_{\boldsymbol{o}}) \tag{3.8}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t * \boldsymbol{c}_{t-1} + \boldsymbol{i}_t * \sigma(\boldsymbol{W}_{\boldsymbol{c}}\boldsymbol{x}_t + \boldsymbol{U}_{\boldsymbol{c}}\boldsymbol{h}_{t-1} + \boldsymbol{b}_{\boldsymbol{c}}) \tag{3.9}$$

$$\boldsymbol{h_t} = \boldsymbol{o_t} * \sigma(\boldsymbol{c_t}) \tag{3.10}$$

式中，$\boldsymbol{h}$是隐藏状态，是$\boldsymbol{c}$单元状态，$\boldsymbol{x}$是输入向量，$\boldsymbol{f}$是遗忘门，$\boldsymbol{i}$是输入门，$\boldsymbol{o}$输出门，，$\boldsymbol{W}$、$\boldsymbol{U}$、$\boldsymbol{b}$是训练参数，$*$是按元素乘。

3.5 本章小结

本文使用基于体素的三维卷积和 PCA 方式、PointNet 和 PointNet++这些模型来提取三维点云的空间域特征，接着使用 RNN、GRU、LSTM 这些用于处理序列数据的神经网络来提取人体动作的三维点云的时间域特征，最后使用全连接层来给出分类结果。对于上述不同模型的组合、以及其不同的参数对分类结果造成的影响将在下一章详细介绍。

第四章 模型性能的实验与分析

4.1 数据集介绍

本文使用了 RadHAR 和 Pantomime 两个数据集来测试第三章中介绍的模型在不同数据集上的表现。

4.1.1 RadHAR 数据集

A. Singh 和 S. Sandha 等人在提出 RadHAR 模型时公布了一套使用 FMCW 雷达采集的人体锻炼动作的三维点云数据集[22]，包含拳击、开合跳、深蹲、跳跃、步行这五个类别的共计 5581 分钟的数据，每秒包含 30 帧点云，每个点云平均包含 21 个点，每个点除了 XYZ 三维坐标，还包含到雷达的距离和方位角、径向速度、多普勒桶、雷达信号强度这 5 个特征值。

4.1.2 Pantomime 数据集

S. Palipana 和 D. Salami 等人在提出 Pantomime 模型时公布了一套使用 FMCW 雷达采集的手臂动作的三维点云数据集[17]，类似于交警手势，包含单臂抬举、双臂前推、单臂顺时针摆动等共计 21 种动作类别，并包含空地、办公室、餐厅、工厂等多个场景，从在雷达不同方位角、不同距离采集的数据、以不同的速度采集的数据，共计包含 19661 条数据，每条数据时长 1~3 秒，平均包含 100 帧，但每帧平均仅包含 3.2 个点，每个点仅包含 XYZ 三维坐标，不含其他特征值。

4.2 数据预处理及数据增强方法

为了在所有模型上应用上述数据集，本文使用了一套数据预处理和数据增强的流程，来适应模型和数据集的多样性，结构如图 4.1 所示。

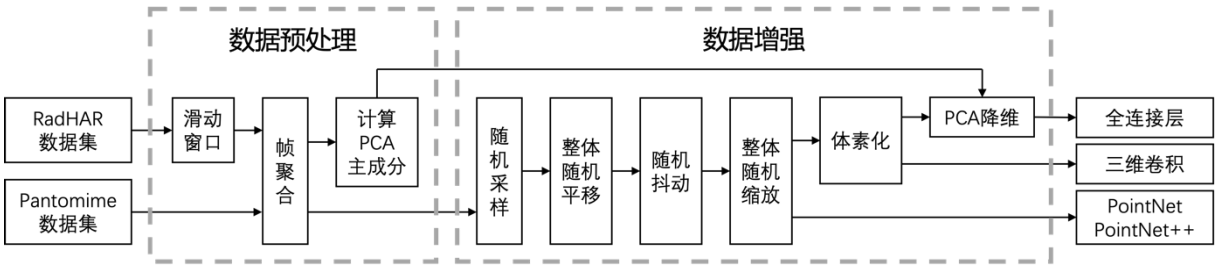


图 4.1 数据预处理和数据增强的流程

4.2.1 帧聚合和滑动窗口

由于数据集每一帧点云的点数都非常少，特征提取层难以从这么少的点中提取出有意义的特征，因此本文聚合了连续的若干帧，使得聚合后的每一帧包含更长时间内的更多点。对于 RadHAR 数据集，本文默认取 10 帧聚合，这样每个聚合帧包含 1/3 秒内的平均 210 个点；对于 Pantomime 数据集，本文默认将原始 1~3 秒的数据平均划分成 4 个聚合帧，这样每个聚合帧包含约 80 个点。

RadHAR 数据集的每条数据包含数十秒的锻炼动作，本文认为锻炼动作时连续且重复的，从其中任意截取 2 秒都是同一类锻炼动作。因此，本文采用了滑动窗口算法对原来数十秒的数据进行截帧，默认窗口大小为 2 秒，步进为 1/3 秒。滑动窗口算法不适用于 Pantomime 数据集，因为 Pantomime 数据集中每条数据仅执行特定动作一遍。

4.2.2 点云采样

由于数据集中每个点云中点的数量是不一致的，为了便于批处理，本文对帧聚合后的点云做随机采样。如果采样点数小于点云点数，本文保证不重复采样；如果采样点数大于点云点数，本文保证每个点至少被采样一次。

4.2.3 平移和缩放变换

为了避免过拟合，本文通过对点云应用随机平移和缩放变换来进行数据增强，主要分为三步进行：

（1） 对所有点的坐标施加符合$N_3(0, 0.1)$分布的随机偏移，来整体移动点云；

（2） 对每一个点独立地施加符合$N_3(0, 0.01)$分布的随机偏移，来对点云的局部特征做数据增强；

（3） 对每个点的坐标施加符合$N(1, 0.1)$分布的随机缩放，来整体缩放点云。

4.2.4 体素化及 PCA 降维

对于体素化的两个模型，本文在训练之前使用全量的未经数据增强的训练集数据来计算 PCA 的主成分，在训练时实时地对经过数据增强的点云数据做体素化，并用 PCA 降维，而非在训练之前先计算好所有体素数据，以此来实现对体素数据的增强。

4.3 实验环境

本章的所有实验在一台 CPU 为 Intel Xeon Gold 5218R，GPU 为 NVIDIA GeForce RTX 3080，内存容量为 64GB、显存容量为 10GB 的服务器上运行。

所有实验代码使用 Python 3.9 编写，基于 PyTorch 1.9.0 机器学习框架，使用的 CUDA 版本为 11.4。

4.4 不同组合的模型性能测试

4.4.1 实验结果

本文在 RadHAR 和 Pantomime 两个数据集上都训练了第三章中介绍的五种点云特征提取层和三种时间序列处理层的所有组合，共 30 组实验，每组实验都进行了调参，选取能达到的最好效果的模型，记录下每个模型收敛之后的测试集准确率、模型可训练参数数量和单样本推理速度，整理如表 4.1 和表 4.2 所示，其中 RadHAR 数据集为 5 分类问题，Pantomime 数据集为 21 分类问题。

表 4.1 RadHAR 数据集上的实验结果

	RNN			GRU			LSTM		
	Acc	Size	Speed	Acc	Size	Speed	Acc	Size	Speed
三维卷积	52.8%	56k	10ms	53.5%	155k	10ms	52.1%	205k	10ms
PCA+多层感知机	56.1%	105k	18ms	66.4%	204k	18ms	60.8%	253k	18ms
PointNet	96.4%	3163k	12ms	95.6%	3820k	12ms	96.3%	4148k	12ms
PointNet++ SSG	96.7%	1064k	20ms	95.6%	1720k	21ms	**97.3%**	2049k	21ms
PointNet++ MSG	95.0%	1280k	35ms	96.4%	1936k	35ms	97.1%	2264k	36ms

表 4.2 Pantomime 数据集上的实验结果

	RNN			GRU			LSTM		
	Acc	Size	Speed	Acc	Size	Speed	Acc	Size	Speed
三维卷积	无法收敛			无法收敛			无法收敛		
PCA+多层感知机	无法收敛			无法收敛			无法收敛		
PointNet	82.8%	3150k	7.9ms	90.7%	3806k	7.9ms	88.3%	4132k	7.8ms
PointNet++ SSG	83.3%	1065k	10ms	92.1%	1721k	10ms	90.1%	2049k	10ms
PointNet++ MSG	91.6%	1281k	17ms	**95.2%**	1937k	17ms	94.6%	2265k	17ms

4.4.2 准确率分析

在 RadHAR 数据集上 PointNet++ SSG 和 LSTM 的组合最高达到了 97.3%的准确率，在 Pantomime 数据集上 PointNet++ MSG 和 GRU 的组合最高达到了 95.2%的准确率。

一个值得注意的现象是，在 RadHAR 数据集上，PointNet、PointNet++ SSG 和 PointNet++ MSG 都基本上达到了 95%以上的准确率，这三者没有明显的区别；而 Pantomime 数据集上，这三个基于 PointNet 的模型的准确率依次提高，最差的 PointNet 的 82.8%与最好的 PointNet++ MSG 的 95.2%构成明显差距。因此，本文认为，Pantomime 数据集是比 RadHAR

数据集更复杂、对模型性能更具有挑战性的，PointNet++ SSG 相比于 PointNet 的改进之处（即对于局部特征的提取能力）能使得它在 Pantomime 数据集上有 2%左右的准确率提升，而 PointNet++ MSG 相比于 SSG 的改进之处（即对于不同点云密度的适应性）带来了另外 4%左右的性能提升。

相比于 PointNet 系列模型，两个基于体素的模型表现得没有那么优秀，在 Pantomime 数据集上甚至于无法收敛，一直维持着接近于随机猜测的准确率。在 RadHAR 数据集上，三维卷积方法达到了 53.5%的准确率，PCA 方法略胜一筹，达到了 66.4%的准确率，但都离 PointNet 系列模型有很大差距，本文认为这是因为大量局部特征在体素化的过程中被丢失了。

对于同一个点云特征提取层，GRU 和 LSTM 的准确率普遍比 RNN 的准确率高，这在 Pantomime 数据集上尤为明显，提升 4%~8%左右。而 GRU 的准确率比 LSTM 略高 2%左右。这也印证了 GRU 和 LSTM 在长期记忆上做出的改进（即更新门、遗忘门等）在本文探究的人体动作三维点云时间序列分类任务上是有必要的。

4.4.3 模型可训练参数数量的分析

基于体素的两种模型的参数数量（50k~200k）普遍比基于 PointNet 的三种模型的参数数量（1000k~4000k）小一两个数量级，这是因为体素方法本身的限制，使得更大的模型不会有更好的效果。

值得注意的是，PointNet++的模型大小仅为 PointNet 的一半，但模型效果却比 PointNet 好 2%左右，这类似于使用卷积神经网络而不是全连接神经网络来处理图像数据，合理地利用平移不变性来共享卷积核参数可以显著降低模型大小的同时还能提升模型的性能。PointNet++的 MSG 方法在 SSG 的基础上增加了约 200k 参数来实现多尺度的局部特征提取，换来的是约 4%的效果提升。

使用 GRU 的模型参数比使用 LSTM 的模型少了约 200k，但效果却提升了约 2%，验证了 GRU 在设计上对 LSTM 结构的简化是合理且有效的。

4.4.4 模型推理时间的分析

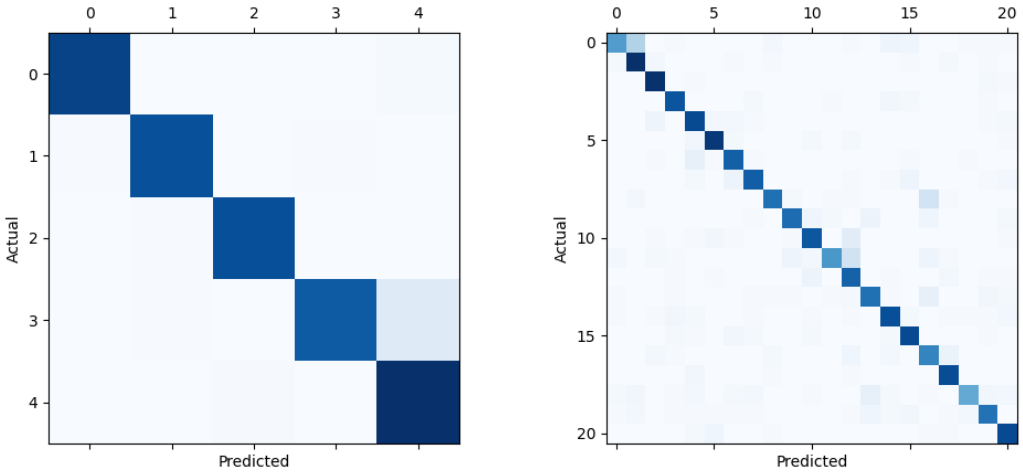
本文测量模型推理时间仅包括了模型前向传播的时间，不包括数据预处理和反向传播的时间。为了充分利用 CPU 和 GPU 的缓存加速，本文测量时跑两边前向传播，第一遍作为模型预热，第二遍才真正计时，使用缓存加速的推理时间仅为不使用的约 1/4。

从测试结果可见，即使是最慢的 PointNet++ MSG，推理速度也仅有 17ms，每秒可推理约 60 次，完全达到了实时处理的性能要求。最快的 PointNet 的推理速度不到 8ms，每秒推理超过 120 次。

虽然 PointNet++的模型大小仅为 PointNet 的一半，但推理速度显著慢于 PointNet，这是因为 PointNet++中共享参数的局部特征提取层需要被计算很多遍。MSG 方法相比于 SSG 方法慢了将近一倍，可见多尺度采样对于速度的影响确实是显著的，因此 Charles R. Qi 等人在 PointNet++论文中提出了多分辨率分组（MRG）方法来解决速度过慢的问题。

4.4.5 混淆矩阵

本文以 PointNet 和 RNN 的组合模型在两个数据集上的分类结果为例，分析其混淆矩阵，如图 4.2 所示。



（a）RadHAR 数据集　　（b）Pantomime 数据集

图 4.2 PointNet 和 RNN 的组合模型在两个数据集上的混淆矩阵

在 RadHAR 数据集上，即使是性能较差的 PointNet 和 RNN 也能达到 96.4%的准确率，唯一出现混淆的是把类别 3（跳跃）识别成类别 4（步行）。在 Pantomime 数据集上，PointNet 和 RNN 的组合就略显不足，错误率较高的有：把类别 0（单臂侧向抬举）识别成类别类别 1（单臂前向抬举），把类别 11（双臂前推）识别成类别 12（双臂回拉）等，这些确实是比较类似的动作。

4.5 数据预处理和数据增强对模型性能的影响

4.5.1 特征通道对模型性能的影响

RadHAR 数据集的每个点除了 XYZ 三维坐标，还提供了雷达的距离和方位角、径向速度、多普勒桶、雷达信号强度这些特征值，本文为了探究这些特征是否会对最终的模型性能产生帮助，进行了多组消融实验。

本文采用了对 RadHAR 数据集效果最好的 PointNet++ SSG 和 LSTM 的组合进行实验，每次实验保留的特征通道和对应的准确率如表 4.3 所示。

表 4.3 保留的特征通道和准确率

距离	方位角	径向速度	多普勒桶	信号强度	准确率
					91.7%
√					**94.6%**
	√				79.2%
		√			**95.7%**
			√		96.5%
				√	91.0%
√		√			**96.7%**
√	√	√	√	√	97.3%

由此可见，距离和径向速度是对于准确率最关键的两个特征。没有任何特征通道（只包含 XYZ 三维坐标）的准确率仅为 91.7%，距离特征使得准确率提升了 2.9%，径向速度特征使得准确率提升了 4.0%，同时加上距离特征和径向速度使得准确率提升了 5.0%，而使用全部的特征比不使用任何特征准确率提升了 6.6%。

另外，距离特征和径向速度特征可以有效提高训练的稳定性。不提供距离和径向速度时，模型在训练过程中在测试集上的准确率有很大范围的波动（75%～90%），提供距离和径向速度特征后，训练时的测试集准确率的波动范围较为稳定（90%～95%）。

出乎意料的是，仅提供方位角特征的模型准确率反而大大降低，只有 79.2%，产生该现象的原因尚不明确。

4.5.2 数据增强对模型性能的影响

数据增强可以有效避免过拟合现象，本文实现了整体随机平移、单点随机抖动、整体随机缩放三种对点云进行数据增强的方法。为了探究这三种数据增强方法对模型结果的影响，本文使用在 Pantomime 数据集上表现最好的 PointNet++ MSG 和 GRU 的模型组合，对三种数据增强方法的随机变量方差做了对比实验，得到的结果如表 4.4 所示。

表 4.4 数据增强对模型性能的影响

整体随机平移距离标准差(m)	单点随机抖动距离标准差(m)	整体随机缩放比例标准差	准确率
0	0	0	90.7%
0.1	0.01	0.1	**95.2%**
0.2	0.02	0.2	90.0%
0.1	0	0	92.1%
0.2	0	0	**94.5%**
0	0.01	0	93.9%
0	0.02	0	**94.5%**
0	0	0.1	**93.6%**
0	0	0.2	89.1%

根据实验结果，适当的数据增强（0.1/0.01/0.1）能够比不数据增强提高 4.5%的准确率，但过度的数据增强（0.2/0.02/0.2）甚至会降低 0.7%的准确率。在三种数据增强方式中，整体随机平移和单点随机抖动的作用比整体随机缩放更大一些，单独施加这两种数据增强也能提高 3.8%的准确率。相比于加上全部三种数据增强方式，单独增加整体随机平移和单点随机抖动方式可以容忍更大的随机距离标准差。

4.6 模型鲁棒性和泛化能力测试

4.6.1 模型在不同背景环境下的表现

该人体动作识别系统是否能在不同的场景下正常使用，是对模型泛化能力好坏的重要指标。本文为了探究模型在未知的场景下能否对人体动作正常识别，限制了训练时使用的训练数据场景范围，在不同场景下的数据进行测试，得到的结果如表 4.5 所示。

表 4.5 不同训练/测试采集场景对模型性能的影响

		训练数据集		
		仅空地	仅办公室	全部
测试数据集	仅空地	**92.3%**	73.6%	93.7%
	仅办公室	81.8%	**90.0%**	94.7%
	仅餐厅	79.4%	68.1%	96.4%
	仅工厂	85.1%	78.3%	97.6%
	空地+办公室	89.5%	85.4%	94.9%
	全部	86.8%	81.9%	**94.2%**

实验的结果并不令人满意，模型对于不同场景的泛化能力还有待提高。在仅空地训练数据集上训练的模型在除了空地的测试数据集上比基准准确率低了平均约 7.7%；在仅办公室训练数据集上训练的模型在除了办公室的测试数据集上比基准准确率低了平均约 12.6%。

4.6.2 模型在不同采集角度下的表现

该系统是否能在不同的采集角度下正常使用，也是判断模型泛化能力好坏的一个指标。本文为了探究模型在非正对人物的角度下能否对人体动作正常识别，限制了训练时使用的训练数据角度，在其他不同角度下的数据进行测试，得到的结果如表 4.6 所示。

表 4.6 不同训练/测试采集角度对模型性能的影响

		训练数据集	
		正对	全部
测试数据集	偏左 45 度	69.9%	88.2%
	偏左 30 度	89.3%	94.1%
	偏左 15 度	92.3%	94.2%
	正对	**94.0%**	96.1%
	偏右 45 度	91.7%	94.3%
	偏右 30 度	77.4%	94.1%
	偏右 15 度	45.2%	82.1%
	全部	92.4%	**95.2%**

对于在全部训练数据集上训练的模型，偏角 30 度以内都能保持和正对相比小于 2%的准确率损失，偏角 45 度的准确率损失约在 10%。对于只在正对数据集上训练的模型，偏角对准确率的损失更加严重一些，相应偏角下比在全部数据集上训练的模型平均低 13%。本文认为偏角太大带来的准确率损失主要应归咎于 FMCW 雷达在非正对的方向上角分辨率下降严重。

4.6.3 模型在不同动作执行速度下的表现

该系统是否能在不同动作执行速度下正常使用，也是判断模型泛化能力好坏的一个指标。本文为了探究模型在人物做出动作的快慢对模型性能的影响，限制了训练时使用的训练数据，在其他执行速度下采集的数据进行测试，得到的结果如表 4.7 所示。

表 4.7 不同训练/测试采集角度对模型性能的影响

		训练数据集	
		正常	全部
测试数据集	慢速	89.3%	91.7%
	正常	**95.3%**	95.3%
	快速	90.5%	94.1%
	全部	92.4%	**95.2%**

本文认为模型在速度上的泛化能力还可以接受。仅在正常速度数据集上训练的模型遇见慢速和快速执行的动作时，准确率和正常速度执行的动作相比平均下降 5.4%，和在全部数据集上训练的模型相应速度的准确率相比平均下降 3.0%。

4.7 本章小结

本章介绍了实验使用的数据集及数据预处理、数据增强的算法流程，比较了不同组合的模型在不同数据集上的共 30 组准确率、模型大小和推理速度，在 RadHAR 数据集和 Pantomime 数据集上分别达到了最高 97.3%和 95.2%的准确率。

本章还对模型进行了大量对比实验，探究了特征通道、数据增强对模型性能的影响，以及模型在不同背景环境、不同采集角度、不同动作执行速度下的表现，得出的结论有：点云距离和速度特征对模型准确率各自能提升 3%左右；适当的数据增强可以提高模型准确率达 4.5%；该模型对于 30 度以内不同偏角的泛化能力和不同动作执行速度的泛化能力比较令人满意，但在不同背景环境下的泛化能力有待提高。

第五章 总结与展望

5.1 研究总结

本文的研究内容可以总结为三个部分：

首先，根据 FMCW 毫米波雷达的工作原理，本文通过距离 FFT、去除静态物体、计算距离-方位角热力图、热力图去噪、去除离群点、多普勒 FFT 这六个步骤，实现了使用 FMCW 雷达给出的原始数据生成三维点云，点云中每个点包含距离、方位角、俯仰角、雷达信号强度、径向速度这五个用于描述物体在三维空间信息的数据。

接着，本文为了对人体动作的三维点云时间序列进行识别和分类，提出了一套由点云特征提取层、时间序列处理层和全连接分类层组成的机器学习模型架构，每个部分可以独立变换组合。在点云特征提取层中使用基于体素的三维卷积和 PCA 方式、PointNet 和 PointNet++这些模型来提取三维点云的空间域特征；在时间序列处理层中使用 RNN、GRU、LSTM 这些用于处理序列数据的神经网络来提取人体动作的三维点云的时间域特征；在分类层中使用全连接层来给出分类结果。

最后，本文比较了不同组合的模型在不同数据集上的共 30 组准确率、模型大小和推理速度，在 RadHAR 数据集和 Pantomime 数据集上分别达到了最高 97.3%和 95.2%的准确率。并对模型进行了大量对比实验，探究了特征通道、数据增强对模型性能的影响，以及模型在不同背景环境、不同采集角度、不同动作执行速度下的表现，得出的结论有：点云距离和速度特征对模型准确率各自能提升 3%左右；适当的数据增强可以提高模型准确率达 4.5%；该模型对于 30 度以内不同偏角的泛化能力和不同动作执行速度的泛化能力比较令人满意，但在不同背景环境下的泛化能力有待提高。

5.2 研究展望

在体素化方面，模型效果对体素粒度和范围非常敏感，导致该方法难以应用到部分数据集，未来值得思考如何更高效地确定体素化的参数，来提高基于体素化模型的性能表现。

在数据集利用率方面，本文的实验发现模型没有很好地利用点云的方位角和雷达信号强度这两个包含有效信息的特征通道，如果能对模型进行改进使得提高模型在这些特征通道上的数据集利用率，可能有助于进一步提升模型在复杂数据集上的性能表现。

在模型泛化能力方面，本文的模型对在不同场景、不同角度、不同动作执行速度下采集的数据的泛化能力都还有进一步提高的空间，在不同场景下的泛化能力尤为欠缺。如果能对泛化能力进行改善，将有助于模型在实际各种应用场景中的综合准确率。

参考文献

[1] 孟岩. 基于 KinectV2 的手势识别技术研究[D]. 燕山大学, 2019.

[2] Shaotran E, Cruz J J, Reddi V J. Gesture Learning For Self-Driving Cars[C]//2021 IEEE Internati onal Conference on Autonomous Systems (ICAS). IEEE, 2021: 1-5.

[3] Attal F, Mohammed S, Dedabrishvili M, et al. Physical human activity recognition using wearable sensors[J]. Sensors, 2015, 15(12): 31314-31338.

[4] Shen C, Ho B J, Srivastava M. Milift: Efficient smartwatch-based workout tracking using automati c segmentation[J]. IEEE Transactions on Mobile Computing, 2017, 17(7): 1609-1622.

[5] Hu J, Lewis F L, Gan O P, et al. Discrete-event shop-floor monitoring system in RFID-enabled manufacturing[J]. IEEE Transactions on Industrial Electronics, 2014, 61(12): 7083-7091.

[6] Aggarwal J K, Ryoo M S. Human activity analysis: A review[J]. Acm Computing Surveys (Csur), 2011, 43(3): 1-43.

[7] Lara O D, Labrador M A. A survey on human activity recognition using wearable sensors[J]. IEE E communications surveys & tutorials, 2012, 15(3): 1192-1209.

[8] 曾上. 基于毫米波的人体动作识别技术研究[D]. 南京大学, 2021.

[9] 徐迎阳. 可穿戴设备现状分析及应对策略[J]. 现代电信科技, 2014 (4): 73-76.

[10] 王星. 基于高分辨率雷达的动态手势识别方法研究[D]. 电子科技大学, 2021.

[11] Lun R, Zhao W. A survey of applications and human motion recognition with microsoft kinect[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2015, 29(05): 1555008.

[12] Lien J, Gillian N, Karagozler M E, et al. Soli: Ubiquitous gesture sensing with millimeter wave radar[J]. ACM Transactions on Graphics (TOG), 2016, 35(4): 1-19.

[13] Wang S, Song J, Lien J, et al. Interacting with soli: Exploring fine-grained dynamic gesture reco gnition in the radio-frequency spectrum[C]//Proceedings of the 29th Annual Symposium on User I nterface Software and Technology. 2016: 851-860.

[14] Li T, Fan L, Zhao M, et al. Making the invisible visible: Action recognition through walls and occlusions[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 872-881.

[15] Zhao M, Liu Y, Raghu A, et al. Through-wall human mesh recovery using radio signals[C]//Proc eedings of the IEEE/CVF International Conference on Computer Vision. 2019: 10113-10122.

[16] Ma Y, Zhou G, Wang S, et al. Signfi: Sign language recognition using wifi[J]. Proceedings of th e ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2018, 2(1): 1-21.

[17] Palipana S, Salami D, Leiva L A, et al. Pantomime: Mid-air gesture recognition with sparse mill imeter-wave radar point clouds[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2021, 5(1): 1-27.

[18] Iovescu C, Rao S. The fundamentals of millimeter wave sensors[J]. Texas Instruments, 2017: 1-8.

[19] Ge L, Liang H, Yuan J, et al. Robust 3d hand pose estimation in single depth images: from sin gle-view cnn to multi-view cnns[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 3593-3601.

[20] Su H, Maji S, Kalogerakis E, et al. Multi-view convolutional neural networks for 3d shape reco gnition[C]//Proceedings of the IEEE international conference on computer vision. 2015: 945-953.

[21] Owoyemi J, Hashimoto K. Spatiotemporal learning of dynamic gestures from 3d point cloud data [C]//2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018: 5929-5934.

[22] Singh A D, Sandha S S, Garcia L, et al. Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar[C]//Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems. 2019: 51-56.

[23] Zhao P, Lu C X, Wang J, et al. mid: Tracking and identifying people with millimeter wave radar[C]//2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS). IEEE, 2019: 33-40.

[24] Qi C R, Su H, Mo K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 652-660.

[25] Qi C R, Yi L, Su H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space[J]. Advances in neural information processing systems, 2017, 30.

[26] Rao S. Introduction to mmWave sensing: FMCW radars[J]. Texas Instruments (TI) mmWave Training Series, 2017.

[27] Stoica P, Wang Z, Li J. Robust capon beamforming[C]//Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, 2002. IEEE, 2002, 1: 876-880.

[28] Gambi E, Ciattaglia G, De Santis A, et al. Millimeter wave radar data of people walking[J]. Data in brief, 2020, 31: 105996.

[29] Farina A, Studer F A. A review of CFAR detection techniques in radar systems[J]. Microwave Journal, 1986, 29: 115.

[30] Schubert E, Sander J, Ester M, et al. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN[J]. ACM Transactions on Database Systems (TODS), 2017, 42(3): 1-21.

[31] Dey R, Salem F M. Gate-variants of gated recurrent unit (GRU) neural networks[C]//2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS). IEEE, 2017: 1597-1600.

[32] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network[J]. Physica D: Nonlinear Phenomena, 2020, 404: 132306.

学位研究期间取得的主要成果

一、 已发表或已录用的论文

无

二、 已获专利或软件著作权

专利类型：计算机软件著作权

专利名称：基于深度学习的多媒体伪造监测取证软件

专利号：2020SR1516610

本人排名：1

著作权人：董世晨、马嘉成，翁静

三、 曾获相关学科竞赛成绩

1. 一级乙等 2021 团体程序设计天梯赛 董世晨 个人二等奖 2021 年 5 月
2. 一级乙等 2020 团体程序设计天梯赛 董世晨 个人二等奖 2020 年 12 月
3. 一级乙等 第十三届全国大学生信息安全竞赛 董世晨 优胜奖 2020 年 8 月
4. 二级乙等 CCF 大学生计算机系统与程序设计竞赛（华东赛区） 董世晨 铜奖 2020 年 10 月
5. 二级乙等 2020 江苏大学生程序设计大赛 董世晨 铜奖 2020 年 11 月

四、 曾主持或参加的大学生创新创业训练计划

无

致　　谢

感谢张玉书老师给我提供的宝贵建议，同时感谢实验室提供的充足的计算资源，使得我能够有机会对模型进行细致的调参，并进行大量对比实验。

感谢南京大学的史书瑜老师对我的悉心指导，对本文研究方向和技术路线指点迷津。

感谢我的父母的支持，以及我的舍友崔明暄、狄文杰、钟璠霖的陪伴。

感谢南京航空航天大学计算机科学与技术学院对本人四年的培养。