计算机科学
COMPUTER SCIENCE

基于雷达和视觉融合的多模态空中手写体识别

刘威, 许勇, 方娟, 李城, 祝玉军, 方群, 何昕

引用本文

刘威, 许勇, 方娟, 李城, 祝玉军, 方群, 何昕. 基于雷达和视觉融合的多模态空中手写体识别[J]. 计算机科学, 2025, 52(9): 259-268.

LIU Wei, XU Yong, FANG Juan, LI Cheng, ZHU Yujun, FANG Qun, HE Xin. Multimodal Air-writing Gesture Recognition Based on Radar-Vision Fusion [J]. Computer Science, 2025, 52(9): 259-268.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240400143

基于雷达和视觉融合的多模态空中手写体识别

刘 威 许 勇 方 娟 李 城 祝玉军 方 群 何 昕
安徽师范大学计算机与信息学院 安徽 芜湖 241002
(2221012420@ahnu.edu.cn)

摘 要 空中手写体识别是一项前景广阔的人机交互技术。单一传感器挖掘手势特征,如毫米波雷达、相机和Wi-Fi,均难以捕捉完整的手势特征。对此,设计了一种灵活的双流融合网络(Two-Stream Fusion Networks,TFNet)模型。该模型既可以融合空中手写体能量图(Air-writing Energy Images,AEIs)和点云时间序列特征图(Point Cloud Temporal Feature Maps,PTFMs),又能仅以单模态数据作为网络的输入。同时,构建了一种鲁棒可靠的多模态空中手写体识别系统。该系统采用硬触发方式启动和结束多传感器数据采集,分别处理同时间序列内的图像和点云数据,生成AEIs和PTFMs,实现多模态数据时间对齐。经过分支网络,对手势外观和细粒度运动信息进行特征提取,结合自适应加权权重,融合双分支决策结果,避免了多模态中间特征的复杂交互,有效地降低了模型的损失。采集多名实验者空中书写0—9共10个数字的空中手写体数据对模型进行评估,结果表明,所提模型在识别精度方面优于其他基线模型,且具有较强的鲁棒性,在空中手写体识别任务中表现出明显优势,可成为多传感器在空中手写体识别任务中的有效工具。

关键词:毫米波雷达;计算机视觉;深度学习;多模态融合;空中手写体识别
中图分类号 TP391

Multimodal Air-writing Gesture Recognition Based on Radar-Vision Fusion

LIU Wei,XU Yong,FANG Juan,LI Cheng,ZHU Yujun,FANG Qun and HE Xin
School of Computer and Information,Anhui Normal University,Wuhu,Anhui 241002,China

Abstract Air-writing gesture recognition is a promising technology for human-computer interaction. Extracting gesture features with a single sensor,such as mmWave radar,camera,or Wi-Fi,fails to capture the complete gesture characteristics. A flexible Two-Stream Fusion Networks(TFNet) model is designed,capable of fusing Air-writing Energy Images(AEIs) and Point Cloud Temporal Feature Maps(PTFMs),as well as operating with unimodal data input. A robust and reliable multimodal air-writing gesture recognition system is constructed. This system utilizes a hard trigger to start and end multi-sensor data acquisition,processing image and point cloud data within the same time sequence to generate AEIs and PTFMs,achieving temporal alignment of multimodal data. Branch networks are employed to extract features of gesture appearance and fine-grained motion information. Adaptive weighted fusion of the dual-stream decision results is used,avoiding the complex interactions of intermediate multimodal features and effectively reducing model loss. Data of ten air-writing gestures representing digits 0—9 are collected from multiple participants to evaluate the model. The results indicate that the proposed model outperforms other baseline models in recognition accuracy and demonstrates strong robustness. The model shows significant advantages in air-writing gesture recognition tasks,making it an effective tool for multi-sensor air-writing gesture recognition.

Keywords mmWave radar,Computer vision,Deep learning,Multimodal fusion,Air-writing gesture recognition

1 引言

随着无线智能感知技术的快速发展,手势识别凭借其自然、直观的交互方式成为人机交互领域的研究热点,不仅广泛应用于智能驾驶[1]、手语表达[2-3]和远程控制[4]等领域,在医疗健康[5]领域也展现出了巨大潜力。尤其在传染病流行时期,基于无线智能感知的手势识别[6]作为无接触人机交互的典范,成为了实现安全交互的重要工具。无线智能感知减少了接触设备的需求,为人机交互提供更安全、更卫生的方式。

空中手写体识别作为手势识别的一个特定应用,专注于识别空中书写中的字母、数字或符号。这种技术通常利用传感器来捕获手部运动轨迹的细微变化[7],使用机器学习或计算机视觉算法识别此类运动信息,进而转化为可识别的文字或符号。空中手写体识别方法主要有基于相机[8-9]、Wi-

到稿日期:2024-04-22 返修日期:2024-10-23
基金项目:国家自然科学基金(62072004)
This work was supported by the National Natural Science Foundation of China(62072004).
通信作者:许勇(yxull@ahnu.edu.cn)

Fi[10-11]和毫米波雷达[12-13]3种。在处理上述感知任务时，单一传感器通常存在特征获取受限和难以适应复杂场景等问题。基于相机的方法通过采集图像或视频数据，利用计算机视觉和深度学习技术提取手势外观特征，以此实现空中手写体识别。然而，相机受光照和极端环境影响较大，光线不足或强光条件下可能导致图像质量下降，影响空中手写体识别的准确性。此外，现有研究虽已实现相机测距[14]，但其无法捕捉目标详细的空间运动状态。基于Wi-Fi的方法则通过信号数据分析，提取空间运动特征以识别手写体。然而，Wi-Fi信号波长相对较长且范围较短，难以捕捉细粒度运动信息。相较于Wi-Fi信号，毫米波信号具有较短的波长和较高的分辨率，能够更精确地捕捉动作细节，但毫米波信号由于无法提供详细的外观特征，且易受多径反射等因素影响，正常环境下的识别精度不及相机。

随着人机交互领域的不断发展，多模态融合[6,15-16]成为该领域重要的研究方向，其中基于毫米波雷达和相机的多模态融合方案备受关注。相机和雷达可以采集高度互补的特征信息，提供丰富的数据源，避免单模态的局限性，显著增强解决复杂任务的能力。本文设计的TFNet模型，既可以融合相机和毫米波雷达传感器，构建鲁棒可靠的多模态空中手写体识别系统，也可仅以单模态数据作为模型的输入，保护用户在特定环境下的隐私。采用硬触发方式启动和结束多传感器数据采集，分别处理同一时间序列内的图像和点云数据，生成AEIs和PTFMs，实现多模态时间对齐。经过分支网络，对手势外观和细粒度运动信息进行特征提取，结合自适应加权权重，融合双分支决策结果，避免复杂的中间特征交互，有效降低了模型的损失。采集多名实验者空中书写0～9共10个数字的空中手写体数据对模型进行评估，结果表明，融合相机和毫米波雷达数据可提高空中手写体识别的准确性和鲁棒性。

本文的主要贡献如下：

1)设计了一种灵活的TFNet模型，该模型既可将相机数据的外观特征与毫米波雷达数据的细粒度运动模式相结合，获得更强大、更全面的空中手写体特征表示，也可仅以单模态数据作为模型的输入，保护用户在特定环境下的隐私。

2)提出了一种特征对齐的方法，该方法采用硬触发方式启动和结束多传感器数据采集，将同一时间序列内的图像和点云数据分别处理为AEIs和PTFMs，实现多模态时间对齐。

3)采集了多名实验者具有时间同步的相机和毫米波雷达空中手写体数据，与其他方法相比较，结果显示，所提方法在准确性和鲁棒性方面具有明显优势。

本文其余部分组织结构如下：第2章回顾了相关研究工作；第3章详细描述了研究方法，包括多模态数据预处理和多模态融合的空中手写体识别；第4章介绍了实验设置；第5章分析实验并展示了实验结果；最后总结全文。

2 相关工作

本章将讨论基于单模态和多模态的不同识别方法。

2.1 基于单模态的识别方法

多种传感器可用于空中手写体识别，如相机、毫米波雷达、Wi-Fi以及可穿戴设备等。对于单一传感器，本文主要讨论毫米波雷达和相机。

2.1.1 基于毫米波雷达的识别方法

空中手写体识别通常被视为分类任务，系统将测试样本准确分类到预定义的类别中以分析人类动作。近年来，研究人员通过不同方式来处理毫米波雷达信号以实现动作分类。在早期的工作中，Singh等[17]提出了RadHAR框架，利用滑动时间窗口和体素化表示来处理稀疏和非均匀的毫米波雷达点云数据，但体素化容易造成细节信息丢失、引入量化噪声，且需要大量内存和计算资源。后续工作中，Yan等[18]提出了基于毫米波雷达的半监督手势识别系统，并比较了距离多普勒图像(Range Doppler Image，RDI)、距离角图像(Range Angle Image，RAI)、多普勒角图像(Doppler Angle Image，DAI)等不同特征图在手势识别中的有效性。但单一种类的特征图保留的信息有限，无法充分利用毫米波雷达所收集到的多种特征。为了能够充分体现出目标的运动状态，Ahmed等[12]提出了基于多流卷积神经网络(Multistream Convolutional Neural Network，MS-CNN)的空中手写体识别方法。具有多个独立输入层的MS-CNN能够创建一个多维深度学习模型，以距离-时间、速度-时间和角度-时间频谱图作为输入，并在后期将所有特征融合，实现手势分类。由于频谱图包含大量冗余信息，为了减少计算资源的开销，Salami等[13]提出了适用于毫米波雷达点云的消息传递神经网络(Message Passing Neural Network，MPNN)图卷积方法。基于点云的动作分类可降低计算复杂性，但由于毫米波雷达点云稀疏，在某些区域可能缺乏足够的数据点来描述手势特征，从而影响手势准确识别。Zhao等[19]提出了复杂加权可学习预处理模块CubeLearn，替代传统的离散傅里叶变换(Discrete Fourier Transform，DFT)预处理方法，可直接从原始信号中提取特征，并构建端到端深度神经网络用于识别任务。近年来，毫米波雷达感知技术在人类动作识别领域发展迅速，能快速捕捉人体运动信息，且不受光照强度的影响。然而，与相机相比，在正常环境下，毫米波雷达对人类细粒度动作识别的准确率有待进一步提高。

2.1.2 基于相机的识别方法

随着相机的普及和深度学习的迅速发展，基于相机的识别方法在空中手写体识别领域占据主导地位。Sharma等[2]设计了一种具有紧凑表示形式的卷积神经网络(Convolutional Neural Network，CNN)模型，相较于其他架构，该模型参数量更少。Sahoo等[20]利用预训练的CNN模型和分数级融合技术，通过端到端微调来实现手势识别。以上方法在有限的静态手势图像数据集上取得了高识别性能，但空中手写体识别更侧重于手势的动态演变过程。对于动态手势识别的研究，Rastgoo等[21]提出了一种简单但高效的模型，将奇异值分解(Singular Value Decomposition，SVD)应用于估计的3D手部关键点坐标。该方法可以获取到更多判别性特征，但SVD对噪声较为敏感。Simonyan等[22]提出了包含两个独立CNN的双流网络，分别用于捕捉视频中静态帧的外观特征和帧间光流的运动特征，然后将这些特征融合实现动作分类。Feichtenhofer等[23]在双流网络结构的基础上，深入研究了多种特征融合策略的有效性。此类双流融合方法证明了多元信息在复杂动作识别中的重要性。目前，基于视觉的动作识别

方法发展迅速，但它们受限于单一数据源，如视频或图像。相机在强光或昏暗环境中对复杂细粒度动作的处理能力有限，缺乏足够的鲁棒性。因此，考虑用多模态融合的方法，来突破单一传感器的局限性。

2.2　基于多模态的识别方法

多模态融合在空中手写体识别中的优势在于，其能够综合不同传感器提供的信息，从而进行更全面、准确的识别。在早期的融合任务中，有研究人员直接对原始雷达数据和相机数据进行融合。例如 Guo 等[24]采用最小二乘法对齐雷达和相机的空间信息，并借助帧内聚类和帧间跟踪算法，从原始传感器数据中提取有效目标信号。Nobis 等[25]将雷达数据从二维地平面转换为垂直图像平面，通过多尺度融合相机数据和投影稀疏毫米波雷达数据。对于此类感兴趣区域(Region of Interest，ROI)生成策略，有效雷达点的数量会直接影响最终的检测结果，且需要对多个传感器进行准确的空间对齐，可能存在计算资源浪费的问题。在之后的研究中，有研究人员考虑对每个模态的信息经过特征提取再进行拼接融合。例如 Chadwick 等[26]将相机收集的图像数据和毫米波雷达扫描生成的距离-速率图像，经过残差特征提取后拼接融合，实现了有效的目标检测。但是，对多模态信息进行单尺度特征拼接，可能会增加模型过拟合的风险。在最近的研究中，Liu 等[6]提出了一种新型的多模态动态手势识别方法。该方法基于具有 Gram 匹配的双分支融合可变形网络，通过融合 RDI 和 RGB 图像，使用 Gram 匹配作为损失函数，有效地挖掘高维异构信息，保持了毫米波雷达和视觉融合的完整性。Shi 等[15]提出了一种双流 CNN 结构，包括身体部位空间注意力(Body-Part Spatial Attention，BPSA)模块和长短时间关系建模(Long-Short Temporal Relation Modeling，LSTRM)模块，分别用于提取摄像头和雷达数据的步态特征，经过双流多尺度特征空间的融合，提升了对微运动特征的捕捉能力。尽管上述方法在多模态融合策略上展现了各自的优势与不足，但针对空中手写体识别的多模态融合研究尚缺乏系统性探索。该领域缺乏具有时间同步的多模态公共数据集，使得研究者在评估方法性能时面临困难。此外，研究者需要在手势分类任务的复杂性与多模态数据对齐和融合策略的复杂性之间进行权衡。本文通过硬触发方式采集多模态数据，并将其处理为 AEIs 和 PTFMs，实现了简便的多模态手势样本对齐。相较于传统的基于双流网络的多模态特征拼接，本文设计了 TF-Net 模型，利用自适应权重加权双流网络，避免了复杂的中间特征交互，有效地降低了模型的损失。针对日益复杂的手势识别场景，挖掘多模态数据的多元信息具有重要意义。

3　研究方法

本章将介绍研究方法的细节，包括数据预处理和多模态融合的空中手写体识别。

3.1　系统概述

图 1 展示了基于多模态融合的空中手写体识别系统，使用 HIKVISION DS-E14 相机收集光学数据，同时利用 TI IWR1642-BOOST 评估板收集回波信号。通过硬触发方式采集一个完整的手写动作，作为一组时间序列数据。其中，将相机收集的原始光学数据按帧提取为图像数据，并通过手势分割和通道平均融合得到 AEI。将毫米波雷达收集的手势回波信号通过快速傅里叶变换得到点云数据，并将点云映射成三维表示，经排列剪裁投影成一张 PTFM。点云反映了手势部位的空间坐标等特征随时间的变换关系，具体的傅里叶变换过程将在 3.2 节中展示。通过 TFNet 模型，从对齐的 AEIs 和 PTFMs 中提取手势特征，结合自适应加权权重，对分支网络的决策结果进行加权融合。该模型避免了多模态中间特征的复杂交互，在空中手写体识别任务中表现出了明显优势。

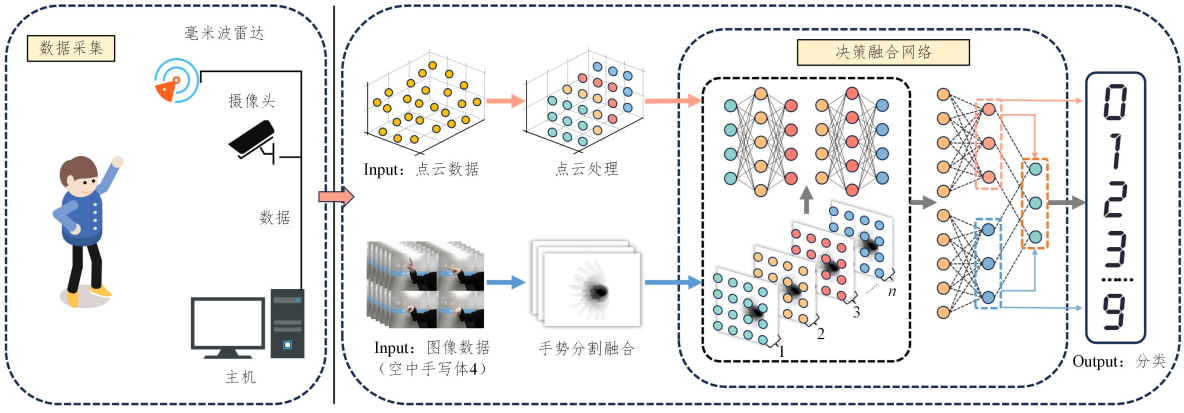


图 1　基于多模态融合的空中手写体识别系统

Fig. 1　Air-writing gesture recognition system based on multimodal fusion

3.2　多模态数据预处理

本节主要介绍对多模态数据的预处理过程，包括 AEIs 和 PTFMs，为后续多模态融合做准备。

3.2.1　空中手写体能量图(AEIs)

运用基于 YCrCb 颜色空间和椭圆形状的肤色分割方法，可以有效地分割图像中的手势区域。首先，利用 YCrCb 颜色空间的特性，其中 Cr 和 Cb 分量代表图像的色度信息。在这种颜色空间下，肤色区域呈现出强烈的聚类特性。将 RGB 颜色空间转换为 YCrCb 颜色空间：

$$\begin{bmatrix} Y \\ Cr \\ Cb \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.331 & 0.5 \\ 0.5 & -0.419 & -0.081 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{1}$$

其次，为了精准地分割肤色区域，采用了椭圆模型对皮肤区域进行建模，使用式(2)描述椭圆：

$$\frac{(Cr-Cr_{center})^2}{a^2}+\frac{(Cb-Cb_{center})^2}{b^2}\leqslant 1 \tag{2}$$

其中，$(Cr_{center}, Cb_{center})$代表椭圆的中心坐标，$a$ 和 b 分别表示椭圆的长轴和短轴。利用预先设定的阈值和椭圆参数，能够有效地将肤色区域与其他区域分离，实现图像的肤色分割。

最后，取一组时间序列图像，对这组数据进行通道平均融合，生成 AEI，$AEI\in\mathbb{R}^{H\times W\times C}$。其中，$H$ 表示 AEI 的高度，W 表示宽度，C 表示通道数。AEI 生成流程如图 2 所示。

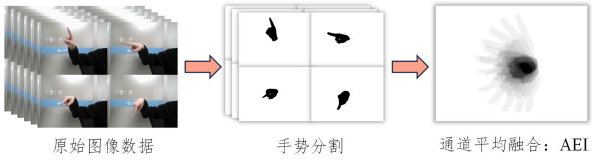


图 2 AEI 生成流程

Fig. 2 AEI generation process

3.2.2 点云时间序列特征图(PTFMs)

在处理毫米波雷达信号时，首先进行第一次傅里叶变换，将时域信号转换成频域信号。毫米波雷达的合成器生成调频信号(Chirps)，并通过 TX 天线发射传输，同时将数据发送至混频器。当信号与目标物体相互作用并发生反射时，RX 天线接收回波信号，并在混频器中产生中频(Intermediate Frequency，IF)信号。IF 的频率 f_0 与雷达到目标的距离 d 的关系表示为：

$$f_0=S\cdot\tau=\frac{S\cdot 2d}{c}\Rightarrow d=\frac{cf_0}{2S} \tag{3}$$

其中，S 表示毫米波信号在频率-时间图上的斜率，即毫米波信号的频率随着时间线性增加；c 代表毫米波信号的速度；τ 代表从 TX 发出信号经反射后到达 RX 所经历的时间。

利用多普勒效应进行第二次傅里叶变换，用于估计目标物体的速度。毫米波雷达一帧发送多个连续的 chirps，物体的移动速度 v 可以从两个相邻的 chirps 之间的多普勒效应引起的相位差 $\Delta\varphi$ 计算得到：

$$\Delta\varphi=\frac{4\pi vT_c}{\lambda}\Rightarrow v=\frac{\lambda\Delta\varphi}{4\pi T_c} \tag{4}$$

其中，λ 为毫米波信号的波长，T_c 代表两个相邻的 chirps 之间的时间间隔。

第三次傅里叶变换通过分析相邻 RX 天线之间的相位变化，计算出目标物体的方向或角度。相位差 $\Delta\varphi$ 和到达角度 θ 之间的关系可以表示为：

$$\Delta\varphi=\frac{2\pi l\sin\theta}{\lambda} \tag{5}$$

其中，l 表示两个 RX 天线间的距离。

通过以上流程生成的毫米波雷达点云数据如图 3 所示。对一组时间序列内的点云数据进行排列剪裁，映射成三维表示，即 PTFM，$PTFM\in\mathbb{R}^{F\times N\times P}$。$F$ 表示序列内的总帧数，N 表示每帧包含的点数，P 表示每个点包含的特征。毫米波雷达经过滤波去噪后，每帧采集的点数具有不均匀性，故取中间值 30。对于少于 30 个点的帧，使用零值进行填充；对于超过 30 个点的帧，则进行裁剪。PTFM 的结构如图 4 所示。

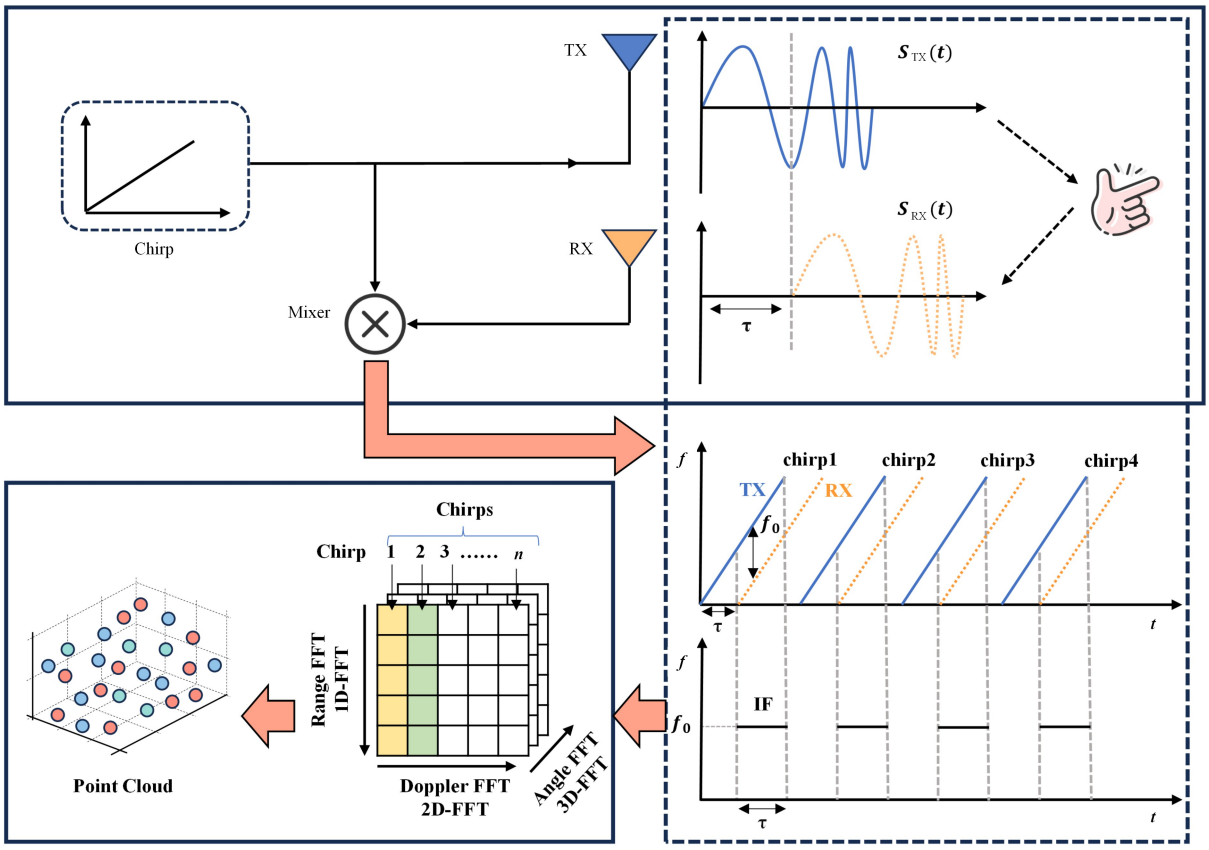


图 3 雷达结构及信号处理流程

Fig. 3 Processing flow of radar structure and signal

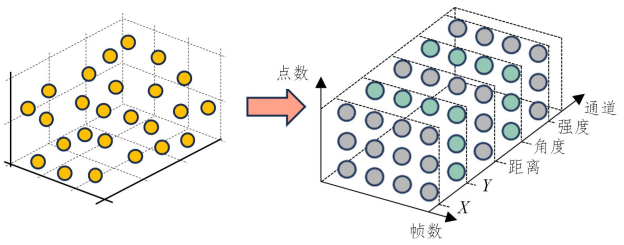


图 4　PTFM 结构

Fig. 4　Structure of PTFM

3.3　TFNet 双流融合网络

本文提出的双流融合网络总体架构如图 5 所示，该架构由双流神经网络组成。在毫米波雷达和相机环境下，将一批 PTFMs 样本和一批 AEIs 样本输入双流网络。首先 PTFMs 通过局部特征提取模块（Local Feature Extraction Block，LFB）提取低水平的手势特征 $\boldsymbol{F}_{\mathrm{PL}}$。随后，通过 LFB 模块提取高水平手势特征 $\boldsymbol{F}_{\mathrm{PH}}$。低水平特征包含了底层的细节，而高水平特征能捕捉到语义层次的信息。对得到的 $\boldsymbol{F}_{\mathrm{PH}}$ 经零填充后，与 $\boldsymbol{F}_{\mathrm{PL}}$ 进行特征拼接（Concatenate）。这种层次化的特征提取和连接方式可以同时利用两个不同尺度的特征进行学习，有助于提高模型对特征的提取能力和学习能力。特征拼接具体描述如下：

$$\boldsymbol{F}_{\mathrm{C}}=\boldsymbol{F}_{\mathrm{PL}}\parallel\boldsymbol{F}_{\mathrm{PH}} \tag{6}$$

其中，$\parallel$ 表示拼接操作，$\boldsymbol{F}_{\mathrm{C}}$ 表示拼接后的张量。接下来，对 $\boldsymbol{F}_{\mathrm{C}}$ 进行最大池化（Max Pooling，MP）操作，提取重要的特征信息并减少模型的计算负担，降低过拟合风险。随后，引入高效通道注意力[27]（Efficient Channel Attention，ECA）学习通道之间的相关性，自适应地调整通道的权重，提高模型的适应性。最后，将得到的张量展平，并经过全连接（Fully Connected，FC）处理，捕捉不同特征间的复杂关系，输出张量表示为 $\boldsymbol{F}_{\mathrm{PF}}$。

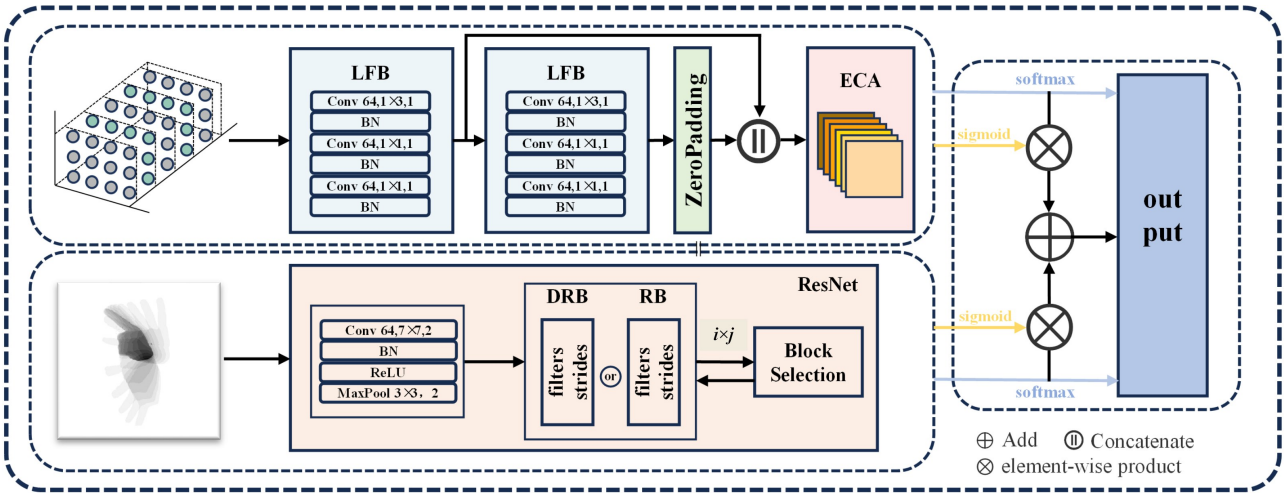


图 5　双流融合网络

Fig. 5　Two-stream fusion network

AEIs 首先经过初始的卷积层和批归一化层，然后执行步长为 2 的 MP 操作，对输入的图像进行空间降维，加快网络收敛速度。接着，经过残差空间深度挖掘语义信息，避免梯度消失等问题。最后，对张量进行全局平均池化（Global Average Pooling，GAP）和 FC 处理，使模型更好地理解整体特征，输出张量表示为 $\boldsymbol{F}_{\mathrm{AF}}$。

将提取到的张量 $\boldsymbol{F}_{\mathrm{PF}}$ 和 $\boldsymbol{F}_{\mathrm{AF}}$，利用 Sigmoid 激活函数和 Softmax 激活函数，得到各自的加权概率分布，之后进行逐元素相加操作，得到融合概率分布。具体方法如下：

$$P_{\mathrm{fusion}}=(f(\boldsymbol{F}_{\mathrm{PF}})\otimes g(\boldsymbol{F}_{\mathrm{PF}}))\oplus(f(\boldsymbol{F}_{\mathrm{AF}})\otimes g(\boldsymbol{F}_{\mathrm{AF}})) \tag{7}$$

其中，P_{fusion} 表示融合后的概率分布，f 表示 Sigmoid 激活函数，g 表示 Softmax 激活函数，$\otimes$ 表示逐元素相乘操作（Multiply），$\oplus$ 表示逐元素相加（Add）操作，经此操作可对 P_{fusion} 输出分类。

若以单模态数据作为输入，则选择对应的分支网络对 $\boldsymbol{F}_{\mathrm{PF}}$ 或 $\boldsymbol{F}_{\mathrm{AF}}$ 进行 FC 操作，利用 Softmax 激活函数直接输出分类。

3.3.1　LFB 模块

LFB 模块由 3 组卷积层与批归一化层组成，每一组的输出 out 可以表示为：

$$out=BN(\sigma(Conv(\boldsymbol{x}))) \tag{8}$$

其中，$\boldsymbol{x}$ 表示输入的张量，σ 表示 ReLU 激活函数，BN 表示批归一化操作，$Conv$ 表示卷积操作。批归一化操作可以在一定程度上加快收敛速度，减少梯度消失或梯度爆炸的情况。

3.3.2　ECA 模块

ECA 模块结构如图 6 所示，使用了一种不降维的局部跨通道交互策略，通过自适应卷积核大小和 GAP 操作，在保持计算效率的同时，增强了模型对重要特征通道的关注。首先自适应确定卷积核大小：

$$k=\psi(C)=\frac{\log_2(C)+b}{\gamma} \tag{9}$$

其中，C 为输入通道数，b 和 γ 为模块的可调参数。若计算得到的卷积核大小为奇数，则直接使用；若为偶数，则将其调整为下一个奇数。

接着，对输入特征执行 GAP 操作，得到一个通道维度上的全局信息。通过一维卷积操作结合 Sigmoid 激活函数计算通道注意力权重向量，表示为 $\boldsymbol{V}\in\mathbb{R}^{1\times1\times c}$。将计算得到的通道注意力权重与输入特征逐元素相乘，得到基于通道注意力

增强的特征表示：

$$\boldsymbol{Y}=\boldsymbol{V}\otimes\boldsymbol{X} \tag{10}$$

其中，Y 表示基于通道注意力增强的特征，X 表示输入特征。

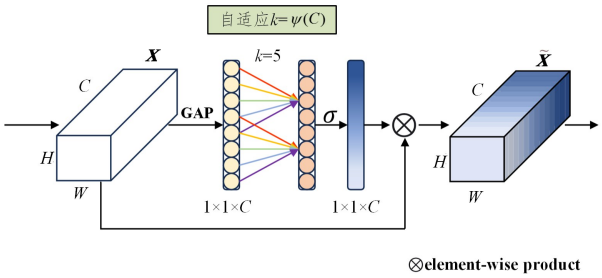


图 6 ECA 模块结构

Fig. 6 Structure of ECA module

3.3.3 ResNet 模块

ResNet 模块结构如图 7 所示，空间第一层残差块的卷积核为 64，之后每层卷积核是上一层的两倍。一个残差块(Residual Block，RB)可以表示为：

$$F=BN(Conv_2(\sigma(BN(Conv_1(\boldsymbol{x}))))) \tag{11}$$

$$y=\sigma(F(\boldsymbol{x},\{Conv_i\})+\boldsymbol{x}) \tag{12}$$

其中，$\boldsymbol{x}$ 代表输入，F 代表残差函数，y 代表残差块的输出，$Conv_i$ 代表第 i 次卷积操作。

最后，判断输入输出维度是否一致，如果一致，则构成了残差块。如果不一致，再使用恒等映射调整维度，即构成了降采样残差块(Downsampling Residual Block，DRB)，表示如下：

$$y=\sigma(F(\boldsymbol{x},\{Conv_i\})+h(\boldsymbol{x})) \tag{13}$$

其中，h 是 1×1 卷积，用于线性调整输入 $\boldsymbol{x}$ 的维度。

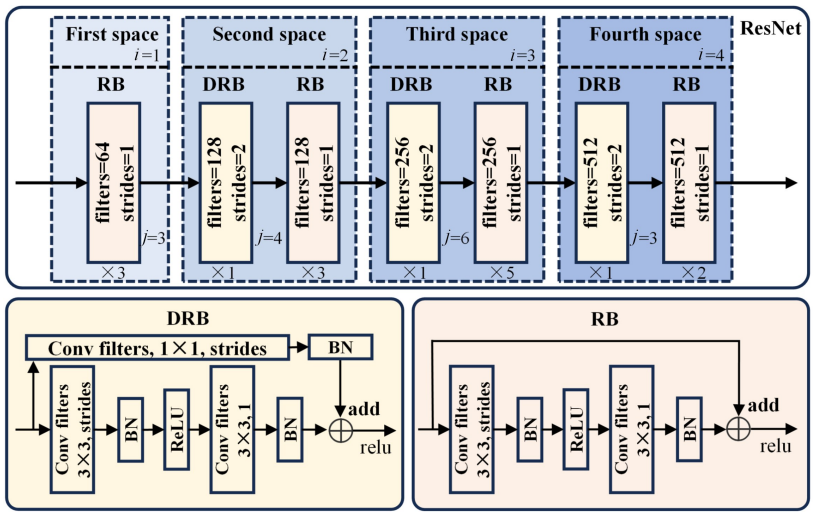


图 7 ResNet 模块结构

Fig. 7 Structure of ResNet module

4 实验设置

4.1 数据采集

在机器人操作系统(Robot Operating System，ROS)环境下，采用硬触发方式同时启动和停止毫米波雷达与相机数据的采集。毫米波雷达和相机数据通过 ROS 消息传送至主机，并使用 ROS 数据包(Rosbag)进行记录和存储。rosbag 中的毫米波雷达数据被转换为文本文件，相机数据被转换为 JPEG 文件。毫米波雷达和相机的采样率分别为 30 FPS 和 10 FPS。图 8 展示了数据采集设备和可视化界面。本实验中设置的毫米波雷达参数如表 1 所列。

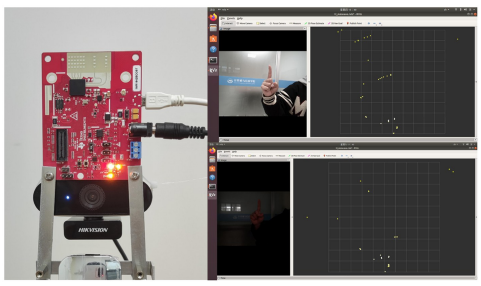


图 8 数据采集设备及可视化界面

Fig. 8 Data collection equipment and visualization interface

表 1 雷达参数设置

Table 1 Radar parameter settings

参数	值
调频起始频率/GHz	77
ADC 采样点数	256
帧数	30
每帧脉冲数	16
帧周期 /ms	33.333

4.2 数据集

本文使用了自建的多模态数据集 Air-Writing 和两种公开的毫米波雷达数据集，下面分别进行介绍。

4.2.1 自建 Air-Writing 数据集

自建 Air-Writing 数据集包含 0～9 共 10 个数字的空中手写体。图 9 展示了 10 种手写体的动作轨迹和实验环境。为了研究不同光照强度对手写体动作分类效果的影响，收集了 12 名实验者在两种光照环境下完成的 10 种空中手写体动作数据。实验者分别在正常办公室环境(Normal Office Environment，NOE)和昏暗办公室环境(Dim Office Environment，DOE)下，对每种动作进行 70 次重复采集。每个动作的持续时间为 2 s。整个数据集共包含 16 800 个点云样本和 16 800 组手势时序图像。对手势时序图像进行手势分割和通道平均融合处理后，得到对应的 16 800 张 AEIs。数据集的具体构成

如表 2 所列。在实验中,毫米波雷达和相机固定于三脚架,与地面的距离约为 1m。实验环境中允许静态物体摆放,除了实验对象之外,无其他运动目标的干扰。毫米波雷达传感器采集到的回波数据经过 3 次傅里叶变换后,得到的点云数据包含空间坐标、距离(点与毫米波雷达的距离)、强度和方位角。

图 9　10 种手写体动作轨迹和实验环境

Fig. 9　Trajectories of ten types of air-writing gestures and experimental environment

表 2　Air-Writing 数据集

Table 2　Air-Writing dataset

手写动作	点云样本		AEIs 样本		总用时/s
	NOE	DOE	NOE	DOE	
0	840	840	840	840	3360
1	840	840	840	840	3360
2	840	840	840	840	3360
3	840	840	840	840	3360
4	840	840	840	840	3360
5	840	840	840	840	3360
6	840	840	840	840	3360
7	840	840	840	840	3360
8	840	840	840	840	3360
9	840	840	840	840	3360

4.2.2　HGR 数据集

Grobelny 等[28]发布了公开的数据集,简称为 HGR 数据集。该数据集使用 AWR1642 雷达传感器收集了 4 个用户的 12 种不同的手部手势。这些手势包括:1)手臂向左(手臂从右向左完全滑动),2)手臂向右(手臂从左向右完全滑动),3)手移开(将一只手从雷达上移开),4)手靠近(将一只手靠近雷达),5)手臂向上(手臂从下到上移动),6)手臂向下(手臂从上到下移动),7)手掌向上(向上旋转手掌),8)手掌向下(向下旋转手掌),9)手向左移动(手向左移动,手臂不动),10)手向右移动(手向右移动,手臂不动),11)水平握拳,12)垂直握拳。雷达收集的点云数据包含距离、速度、峰值、空间坐标。共采集 4600 个样本。

4.2.3　MMActivity 数据集

MMActivity 数据集[17]使用德州仪器 IWR1443-BOOST 雷达来收集两个用户在雷达前执行 5 种不同活动的数据。这些活动包括拳击、直立跳、开合跳、蹲起和原地踏步。雷达的采样率为 30FPS,收集的点云信息包含空间坐标、速度、距离、强度和方位角。共采集 15635 个样本数据。

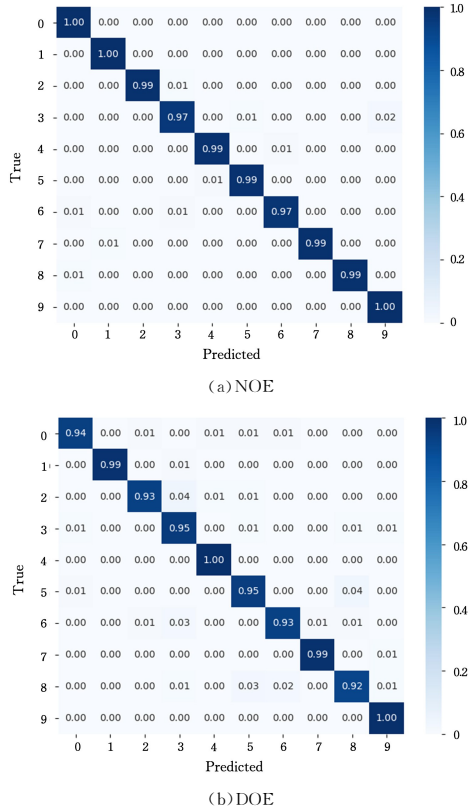
5　实验及分析

5.1　TFNet 性能测试

本节展示了 TFNet 模型在不同光照环境下对 Air-Writing数据集的分类准确率以及对噪声数据进行的鲁棒性测试。

5.1.1　TFNet 模型的测试精度

图 10 展示了一次实验的混淆矩阵,可以明显看出,在 NOE 环境下,TFNet 能够高度准确地区分不同的空中手写体动作,识别准确率达到了 98.71%。在 DOE 环境下,识别准确率有所下降,但依然保持着较高水平,达到了 95.89%。手写动作 2,6 和 8 的分类准确率稍低,可能是由于相机无法捕捉到足够的光线,使图像中的手势特征变得模糊,与其他类别的特征产生混淆。经过多次实验,TFNet 在 NOE 和 DOE 环境下识别的平均准确率分别为 98.48%和 95.56%。结果表明,TFNet 模型具有出色的空中手写体识别效果。



(a) NOE

(b) DOE

图 10　TFNet 在 NOE 和 DOE 环境下的识别准确率

Fig. 10　Recognition accuracy of TFNet under NOE and DOE

5.1.2　噪声数据上的性能

模型的鲁棒性可以通过在数据中添加高斯噪声来验证。为探究 TFNet 模型在不同光照环境下的鲁棒性,对 PTFMs 和 AEIs 分别或整体添加高斯噪声,并经过多次实验测试了识别的平均准确率。TFNet 模型在经过高斯噪声处理的数据集上的表现如图 11 所示。

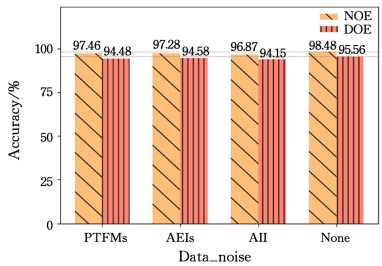


图 11　不同模态数据添加噪声的测试结果

Fig. 11　Testing results of adding noise to different modal data

在 NOE 环境下,向 PTFMs 添加噪声对模型的影响略低

于向 AEIs 添加噪声，在 DOE 环境下结论则相反。这说明在不同环境下 TFNet 模型对两种模态的侧重点不同，也体现了模型对不同模态自适应加权的重要性。此外，在不同环境下，对两种模态数据添加噪声，其识别准确率均略低于对任一模态数据添加噪声。虽然分类准确率略有下降，但依然保持着较高的水平，说明 TFNet 模型仍然能够充分利用两种模态数据间高度互补的信息，弱化噪声的干扰。实验证明了 TFNet 模型具有较强的鲁棒性。

5.2 基线模型与性能测试

本节将介绍所对比的基线模型，并评估单模态和多模态基线模型在自建数据集上的测试精度。

5.2.1 基线模型

为了进一步评估 TFNet 模型的性能，对比了多种常见的单模态和多模态基线模型，包括：

1) LSTM：能够捕捉点云的时间动态特征。

2) ResNet[29]：通过残差连接解决网络的梯度消失问题。

3) DenseNet[30]：提升了特征传递效率和特征提取能力。

4) CNN-LSTM：可以提取点云的空间和时间动态特征。

5) PointNet[31]：能够降低对点云排列方式的敏感性。

6) CNN：通过卷积层和池化层提取图像的空间特征。

7) VGG16[32]：使用多个堆叠的卷积层和小尺寸滤波器，逐层提取图像的高层次特征，广泛应用于视觉任务。

8) InceptionV3[33]：通过混合多尺度卷积核捕捉特征。

9) MobileNet[34]：轻量级的卷积神经网络，通过深度可分离卷积大幅减少参数量，适用于移动和嵌入式设备。

10) Dual-CNN[35]：利用双流 CNN 对多模态手势动作进行特征提取、融合并分类。

11) Fusion-ConvNet[36]：利用空间流和时间流 ConvNet，构建了时空双流融合网络，用于提取动作特征。

这些基线模型涵盖了时间动态特征捕捉、空间特征提取、点云处理以及多模态融合等多种方法。对比这些基线模型可以全面评估 TFNet 模型在多模态融合方面的优势。

5.2.2 基线模型测试结果

为验证多模态融合的有效性，在 NOE 和 DOE 环境下，分别在基线模型上对 Air-Writing 数据集中的毫米波雷达数据和图像数据进行多次测试，并与多模态融合网络 TFNet 进行了比较，平均准确率如表 3 所列。结果显示，相较于基线模型，融合了多模态信息的 TFNet 在 NOE 和 DOE 环境下的识别精度均有所提高。这表明 TFNet 模型能够充分利用两种模态高度互补的特征，提升空中手写体识别的准确率，从而证明了 TFNet 在多模态融合方面的有效性。

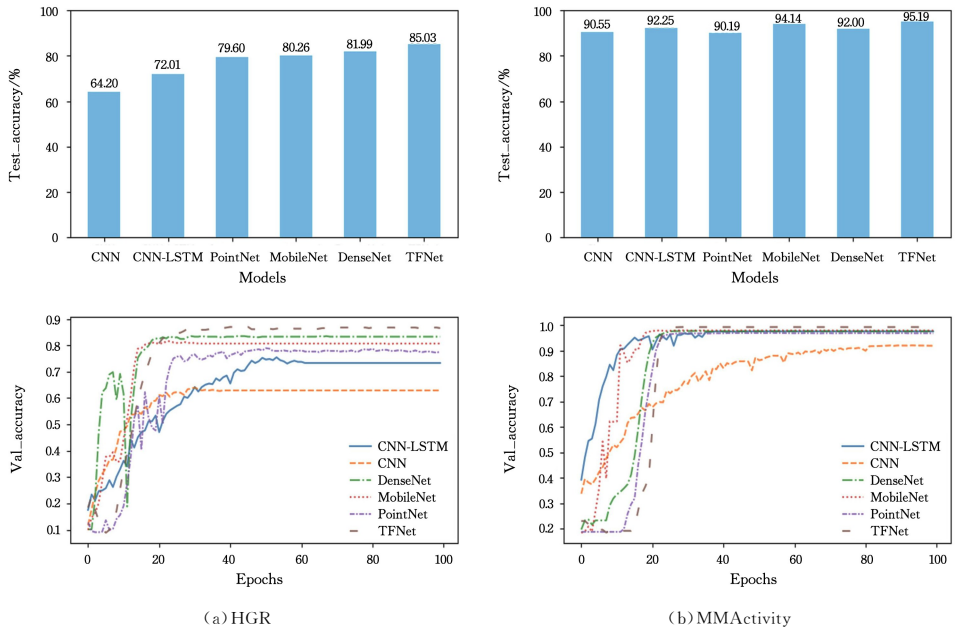
表 3 基线模型测试结果

Table 3 Baseline model testing results

网络	序号	方法	平均准确率/%	
			NOE	DOE
雷达网络	1	LSTM	75.97	75.63
	2	ResNet	79.35	78.65
	3	DenseNet	82.58	81.89
	4	CNN-LSTM	83.62	83.57
	5	PointNet	85.40	85.12
相机网络	1	CNN	93.19	73.51
	2	VGG16	96.71	88.40
	3	InceptionV3	96.74	94.38
	4	MobileNet	97.63	94.52
	5	ResNet	97.84	94.92
融合网络	1	Dual-CNN	97.01	91.73
	2	Fusion-ConvNet	97.36	92.63
	3	TFNet	**98.48**	**95.56**

5.3 在公开数据集上的性能测试

本文使用公开的毫米波雷达数据集 HGR 和 MMActivity，分别对 TFNet 模型中的雷达分支网络和其他基线模型进行了测试，测试集的准确率以及验证集的准确率曲线如图 12 所示。雷达分支网络使用跳跃连接，有助于模型更好地学习特征之间的关联性，提高特征表示能力。同时，引入 ECA 模块自适应地调整每个通道的权重。与基线模型相比，雷达分支网络实现了更高的识别准确率。



(a) HGR (b) MMActivity

图 12 在公开数据集上的识别效果

Fig. 12 Recognition performance on public datasets

5.4　消融实验

本实验通过删除或替换 TFNet 模型中的特定模块，在 Air-Writing 数据集上测试它们对模型性能的影响。在 NOE 和 DOE 环境下，分别对完整的 TFNet 模型进行训练和评估，将其作为基准性能。对输入到分支网络的单模态数据，利用 Softmax 激活函数进行分类，验证了融合的有效性。此外，通过剔除 ECA 模块、将跳跃连接的 LFB 模块替换为卷积神经网络，系统地分析了每个模块对模型性能的影响。结果如表 4 所列。ECA 模块采用了一种不降维的局部跨通道交互策略，通过自适应卷积核大小和 GAP，在保持计算效率的同时，增强了模型对重要特征通道的关注。LFB 模块中的多层卷积操作有助于提取输入的局部特征，批归一化层有助于稳定训练过程。通过局部特征提取和跳跃连接，模型能够更有效地捕获点云的局部信息及其复杂关系。通过对两个分支的自适应加权决策结果进行融合，能够充分挖掘丰富互补的动作特征用于空中手写体识别。因此，使用 LFB 和 ECA 模块以及自适应加权融合多模态数据，有效提升了空中手写体识别的准确率。

表 4　消融实验结果

Table 4　Ablation experiment results

数据输入	网络	LFB	ECA	平均准确率/%	
				NOE	DOE
PTFMs	雷达分支			88.43	87.58
AEIs	相机分支			97.84	94.92
PTFMs AEIs	TFNet	—	—	97.48	94.48
		√	—	98.13	95.04
		√	√	**98.48**	**95.56**

结束语　本文采用硬触发方式启动和结束多传感器数据采集，对齐同一时间序列的多模态数据 PTFMs 和 AEIs，并设计了基于多模态的灵活的双流融合模型 TFNet。在毫米波雷达数据输入时，TFNet 通过局部特征提取模块和跳跃连接，利用高效通道注意力机制赋予通道不同权重，提取手势细粒度运动特征。在相机数据输入时，相机分支网络通过残差模块捕获输入和输出之间的差异，提取具有判别性的手部外观特征。通过结合自适应加权权重，融合双分支决策结果。TFNet 模型不仅可以实现无约束的空中手写体识别，而且与基线模型相比，测试精度明显提高，在 NOE 和 DOE 下识别的平均准确率分别达到了 98.48%和 95.56%。此外，在不同光照条件下对噪声数据的评估结果，也展现了 TFNet 较强的鲁棒性。未来研究可以进一步优化 TFNet 的结构，探索更有效的特征融合策略和数据对齐方法，充分挖掘模态间的多元信息，提高模型的性能。

参 考 文 献

[1] KÖPÜKLÜ O, LEDWON T, RONG Y, et al. Drivermhg: A multi-modal dataset for dynamic recognition of driver micro hand gestures and a real-time recognition framework[C]// 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). IEEE, 2020: 77-84.

[2] SHARMA S, SINGH S. Vision-based hand gesture recognition using deep learning for the interpretation of sign language[J]. Expert Systems with Applications, 2021, 182: 115657.

[3] ZHOU L Y, ZHANG J H, YUAN T T, et al. Sequence-to-Sequence Chinese Continuous Sign Language Recognition and Translation with Multi- layer Attention Mechanism Fusion[J]. Computer Science, 2022, 49(9): 155-161.

[4] LIU H, ZHOU A, DONG Z, et al. M-gesture: Person-independent real-time in-air gesture recognition using commodity millimeter wave radar[J]. IEEE Internet of Things Journal, 2021, 9(5): 3397-3415.

[5] MAHMOUD N M, FOUAD H, SOLIMAN A M. Smart healthcare solutions using the internet of medical things for hand gesture recognition system[J]. Complex & Intelligent Systems, 2021, 7: 1253-1264.

[6] LIU H, LIU Z. A multimodal dynamic hand gesture recognition based on radar-vision fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-15.

[7] TANG X, YAN Z, PENG J, et al. Selective spatiotemporal features learning for dynamic gesture recognition[J]. Expert Systems with Applications, 2021, 169: 114499.

[8] WATANABE T, MANIRUZZAMAN, HASAN M A, et al. 2D Camera-based air-writing recognition using hand pose estimation and hybrid deep learning model[J]. Electronics, 2023, 12(4): 995-1009.

[9] QI J, MA L, CUI Z, et al. Computer vision-based hand gesture recognition for human-robot interaction: a review[J]. Complex & Intelligent Systems, 2024, 10(1): 1581-1606.

[10] LIN C, AHMAD A, QU R, et al. A handwriting recognition system with wifi[J]. IEEE Transactions on Mobile Computing, 2023, 23(4): 3391-3409.

[11] GUO Z, XIAO F, SHENG B, et al. WiReader: Adaptive air handwriting recognition based on commercial WiFi signal[J]. IEEE Internet of Things Journal, 2020, 7(10): 10483-10494.

[12] AHMED S, KIM W, PARK J, et al. Radar-based air-writing gesture recognition using a novel multistream CNN approach[J]. IEEE Internet of Things Journal, 2022, 9(23): 23869-23880.

[13] SALAMI D, HASIBI R, PALIPANA S, et al. Tesla-rapture: A lightweight gesture recognition system from mmwave radar sparse point clouds[J]. IEEE Transactions on Mobile Computing, 2022, 22(8): 4946-4960.

[14] YAN X Y, LU F F, GE L S, et al. Image Style Transfer Based on the Distribution Matching of the Style Features[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2023, 40(3): 48-55.

[15] SHI Y, DU L, CHEN X, et al. Robust gait recognition based on deep CNNs with camera and radar sensor fusion[J]. IEEE Internet of Things Journal, 2023, 10(12): 10817-10832.

[16] CHEN Y S, CHENG K H. BiCLR: Radar-Camera-based Cross-Modal Bi-Contrastive Learning for Human Motion Recognition [J]. IEEE Sensors Journal, 2024, 24(3): 4102-4119.

[17] 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.
[18] YAN B, WANG P, DU L, et al. mmGesture: Semi-supervised gesture recognition system using mmWave radar[J]. Expert Systems with Applications, 2023, 213: 119042.
[19] ZHAO P, LU C X, WANG B, et al. Cubelearn: End-to-end learning for human motion recognition from raw mmwave radar signals[J]. IEEE Internet of Things Journal, 2023, 10(12): 10236-10249.
[20] SAHOO J P, PRAKASH A J, PŁAWIAK P, et al. Real-time hand gesture recognition using fine-tuned convolutional neural network[J]. Sensors, 2022, 22(3): 706-720.
[21] RASTGOO R, KIANI K, ESCALERA S. Real-time isolated hand sign language recognition using deep networks and SVD [J]. Journal of Ambient Intelligence and Humanized Computing, 2022, 13(1): 591-611.
[22] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems, 2014, 27: 568-576.
[23] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1933-1941.
[24] GUO X, DU J, GAO J, et al. Pedestrian detection based on fusion of millimeter wave radar and vision[C]// Proceedings of the 2018 International Conference on Artificial Intelligence and Pattern Recognition. 2018: 38-42.
[25] NOBIS F, GEISSLINGER M, WEBER M, et al. A deep learning-based radar and camera sensor fusion architecture for object detection[C]// 2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF). IEEE, 2019: 1-7.
[26] CHADWICK S, MADDERN W, NEWMAN P. Distant vehicle detection using radar and vision[C]// 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 8311-8317.
[27] WANG Q, WU B, ZHU P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11534-11542.
[28] GROBELNY P, NARBUDOWICZ A. Hand gestures recorded with mm-Wave FMCW radar (AWR1642) [DB/OL]. (2021-06-03) [2024-06-02]. https://dx.doi.org/10.21227/wh5w-c362.
[29] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
[30] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4700-4708.
[31] 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.
[32] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409.1556, 2014.
[33] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2818-2826.
[34] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv: 1704.04861, 2017.
[35] XU L, ZHANG K, YANG G, et al. Gesture recognition using dual-stream CNN based on fusion of sEMG energy kernel phase portrait and IMU amplitude image[J]. Biomedical Signal Processing and Control, 2022, 73: 103364.
[36] CHEN J C, LEE C Y, HUANG P Y, et al. Driver behavior analysis via two-stream deep convolutional neural network[J]. Applied Sciences, 2020, 10(6): 1908-1922.

LIU Wei, born in 2000, postgraduate. His main research interests include wireless intelligent sensing and deep learning.

XU Yong, born in 1966, Ph.D, professor, master supervisor. His main research interests include computer network security, IoT security, wireless intelligent sensing and deep learning.

(责任编辑:何杨)