采用传输控制的多模态数据协作式任务导向通信

Siqi Wan*、Qianqian Yang†、Zhiguo Shi†、Zhaohui Yang† 和 Zhaoyang Zhang†
*浙江大学理工学院，中国杭州。电子邮箱：22260112@zju.edu.cn
†浙江大学信息与电子工程学院，中国杭州。电子邮箱：22260112@zju.edu.cn、qianqianyang20@zju.edu.cn、shizg@zju.edu.cn、yang.zhaohui@kcl.ac.uk、ningming@zju.edu.cn

arXiv:2302.02608v1 [cs.IT] 6 Feb 2023

摘要——物联网（IoT）环境中的实时智能应用依赖及时的数据通信。然而，传输和分析海量多模态数据颇具挑战。近期提出的基于深度学习的面向任务的通信方法已展现出其在通信效率方面的优势。本文提出一种协作式任务导向通信方法，用于将多台终端设备的多模态数据传输至中央服务器。具体而言，我们利用速率较低的一种模态数据的传输结果，控制速率较高的其他模态数据的传输，从而减少传输数据量。我们以智能家居环境中的人体活动识别（HAR）任务为例，为传输不同房间的监控视频及被监测者的加速度数据设计语义导向收发机。数值结果表明，利用基于接收加速度数据所得结果的传输控制，在保持 HAR 任务性能的同时，传输量可降至未采用传输控制时的 2%。

一、引言

随着通信技术的发展以及硬件计算能力的提升，我们正处于万物互联的时代。大量设备接入通信网络，并在任何时间、任何地点产生海量异构数据。智能物联网应用需要低时延、高可靠的数据传输。传统通信方式侧重于比特传输，似乎已不足以应对这一持续增长的需求。受机器学习（ML）技术，尤其是深度学习（DL）进步的推动，语义通信方法在传输效率方面相较传统方法取得了显著提升，正受到越来越多的关注 [1], [2]。

传统通信系统的目标是准确传输比特流，却忽略了语义内容以及对底层任务的有效性 [3]。然而，语义通信或面向任务的通信系统基于深度学习（DL）对语义编码器和解码器进行端到端优化，仅通过信道发送与任务相关的语义特征。针对文本、图像和语音的数据传输，已有研究提出了有效的语义通信系统 [4]–[8]。除这些重建任务外，基于深度学习的语义通信系统也已应用于不同的人工智能任务，例如场景分类 [9] 和目标识别 [10]。

现有的大多数语义通信研究都考虑在一对发射机与接收端之间传输单一模态。然而，许多物联网应用涉及将音频、视频、感知数据等多模态数据从分布式物联网设备传输至中央服务器。文献 [11], [12],提出了基于多模态数据的面向任务的语义通信系统，与传统方法相比，该系统提高了传输效率以及目标任务的性能。混合自动重传请求等传统通信策略也已应用于面向任务的通信系统，以进一步提高通信可靠性 [13],[14]。受这些研究启发，本文提出一种协作式面向任务的通信方法，利用不同模态之间的语义关系，在不降低任务性能的前提下最大限度地降低传输成本。本文的贡献总结如下：

- 我们提出了一种面向任务的通信方法，用于在发送不同模态数据的分布式终端设备与中央服务器之间进行数据传输。我们引入了一种传输控制机制，根据不同模态之间的语义相关性，利用一种模态的传输结果来控制另一种模态的数据传输。
- 我们以人体活动识别（HAR）任务为例，考虑一个由部署在三个不同房间中的三个摄像头、佩戴在受监测人员身上的加速度计以及一台中央服务器组成的智能家居环境。我们设计了一种基于 3D-CNN 的语义收发器，用于传输三个摄像头拍摄的监控视频；并设计了一种基于随机森林的语义收发器，用于传输加速度计采集的加速度数据。根据加速度数据得出的姿态识别结果用于控制视频语义特征的上传。
- 仿真结果表明，通过采用面向语义的传输和数据传输控制机制，所提出的通信框架在不降低 HAR 任务准确率的情况下，大幅降低了通信开销。

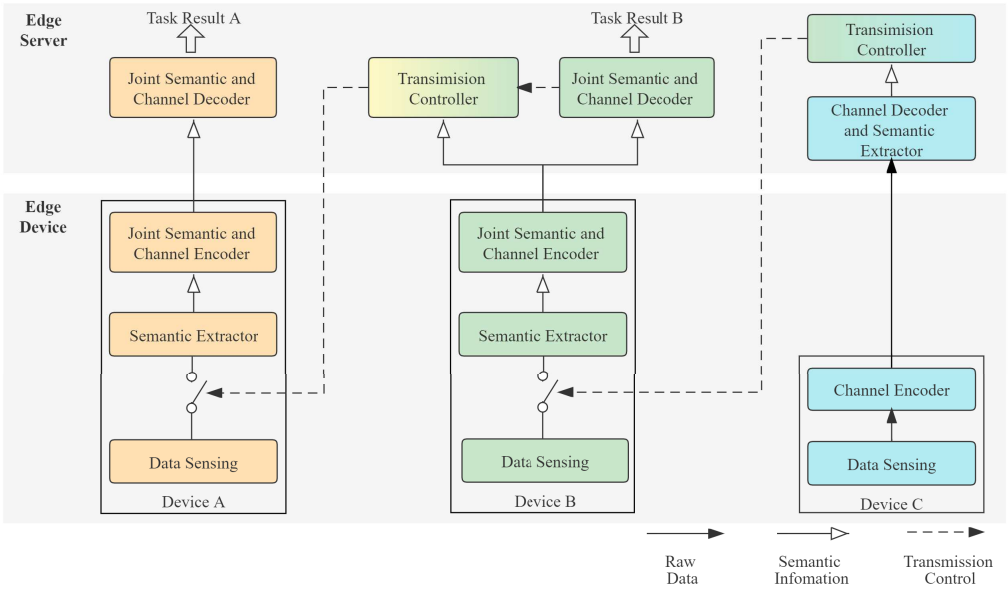

图 1. 基于边云架构和多模态数据融合的面向任务的物联网通信框架系统模型。

本文其余部分安排如下：第二节介绍所提出通信框架的系统模型，并给出智能家居中面向 HAR 任务的语义通信框架示例。第三节详细介绍 HAR 任务的语义通信模型和传输控制机制。第四节展示仿真结果。第五节总结全文。

II. 系统模型

如图 1 所示，我们考虑一种用于多模态数据传输的面向任务的通信系统。系统中有多个设备，每个设备传输不同模态的数据，以共同完成一项任务。该系统由多个发射器和一个接收端组成。发射器是负责采集和传输不同类型数据的边缘设备；接收端则是中央服务器或云服务器，利用接收到的数据执行目标任务，并发送控制信号以触发边缘设备进行传输。

1) 在发射机侧，每个发射机收集不同模态的数据，以进行语义提取和编码，通常由如图 1 所示的三个模块组成。数据感知模块收集感知信号，生成原始数据。语义提取器模块从原始数据中提取语义信息，并利用深度学习方法等语义分析算法移除无关信息。联合语义与信道编码器模块进一步压缩提取出的语义信息，并生成一系列将在物理信道上传输的符号。语义提取器模块和联合语义与信道编码器模块的计算开销较大，因此通常安装在具有充足计算资源的设备上，例如智能摄像头和智能音箱。需要注意的是，一些边缘设备的计算能力非常有限，例如红外测距传感器和温度计等简单传感器。这些设备无法执行复杂的语义提取和编码算法，且通常数据速率较低，例如温度信息。因此，这些设备仅具有数据感知模块和信道编码器模块，直接对感知到的原始数据进行编码并传输至接收端。

2）接收端通过相应的信道与语义联合解码器模块，对来自各边缘设备的接收信号进行信道与语义解码，并输出给定任务的结果。此外，接收端还利用传输控制器模块，根据不同模态之间的语义关系，使用一种模态的语义信息控制另一种模态的数据传输，从而尽量减少传输数据量。需要注意的是，由于计算能力有限，一些边缘设备会直接将原始数据发送至服务器，而不进行语义提取或处理。对于这些接收到的消息，接收端还会通过信道编码器和语义提取器模块重建原始数据，并从中提取语义信息。随后，传输控制器会根据提取出的语义信息，决定其他边缘设备的数据传输。

如图 2 所示，我们针对智能家居中的 HAR任务采用了所提出的面向任务的通信框架。三个不同的房间（卧室、客厅和厨房）分别安装了一台具备一定计算资源的摄像头，用于录制监控视频以监测人体活动；此外，目标人员的左腕下佩戴了一台加速度计，用于获取

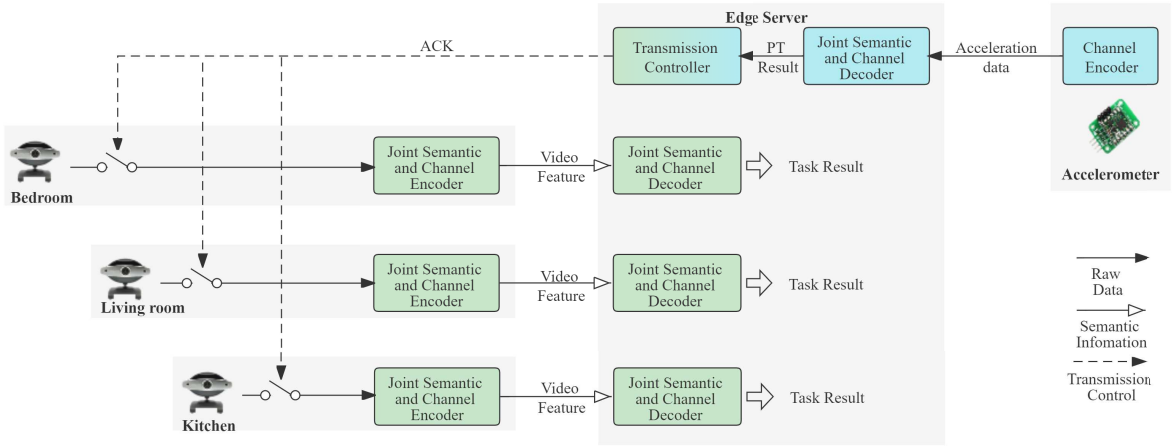

图 2. 面向任务的语义通信网络的结构。

被监测人员的加速度数据。如上所述，每台摄像头均配备联合语义与信道编码器模块，用于从录制的监控视频中提取与人体活动相关的语义特征，并将其编码为传输符号。与此同时，加速度计是一种计算资源有限的小型设备，只能采集并上传原始数据。加速度计中的信道编码器模块会将采集到的加速度数据编码为发送符号，并通过无线信道发送。云服务器接收到这些符号后，联合语义与信道解码器模块会对其进行解码和分析，以获得人体姿态转变（PT）结果。需要指出的是，当人体活动发生变化时，人体姿态必然也会发生变化。考虑到人体姿势与活动之间的这种语义关系，我们在服务器上使用传输控制器，由其输出确认信号，以控制三台摄像头上传人体活动语义特征。只有在摄像头收到肯定的确认信号后，才会从此时起通过本地的联合语义与信道编码器开始上传监控视频。随后，服务器上相应的联合语义与信道解码器模块会对接收到的特征进行解码和分析，生成活动分类结果。下一节将介绍所提出系统的详细内容。

III. 面向 HAR 的通信框架

本节将详细介绍所提出的带传输控制的 HAR 任务导向系统，包括从三个摄像头向服务器传输监控视频的收发器设计，以及从加速度计向服务器传输加速度数据的收发器设计。

A. 用于监控视频的语义收发器

用于监控视频传输的语义收发器如图3所示，其中包括联合语义与信道编码器（位于每个摄像头中）以及接收端对应的联合语义与信道解码器。每个摄像头在每个房间内录制监控视频，视频由连续的图像帧组成。随后，摄像头根据云服务器发出的确认信号，决定是否进一步处理录制的视频。如果摄像头收到肯定的确认信号，则从该时刻起发送一个由16个连续图像帧组成的视频片段，这些图像帧通过滑动窗口采样得到，该视频片段表示为一个四维矩阵 $x_i \in \mathbb{R}^{16\times112\times112\times3}$，其中 i 表示当前样本的索引，最后一个维度表示RGB图像的三个信道。随后，该矩阵 x_i 被输入联合语义与信道编码器，该编码器由一个三维（3D）卷积层和一个最大池化层组成：卷积层使用大小为（3,3,3）的卷积核和（1,1,1）的填充，最大池化层使用（5,5,3）的卷积核及（5,5,3）的步幅，并生成一个视频语义特征矩阵，记为 $k_i \in \mathbb{R}^{4\times5\times22\times22}$ [15]。最后，作为信道编码器的重塑层将高维语义特征映射为语义符号序列，记为 $s_i \in \mathbb{R}^{4840\times2}$，其中第一和第二个信道分别对应复信号的实部和虚部。

符号 s_i 通过物理信道发送至中央服务器。接收到的语义符号记为 $\hat{s}_i = s_i + n$，其中 $\hat{s}_i \in \mathbb{R}^{4840\times2}$ 和 $n \sim N(0, \sigma^2)$ 表示均值为零、方差为 σ 的加性高斯白噪声(AWGN)。物理信道的信噪比 (SNR)（单位：dB）为

$$SNR = 10\log_{10}\left(\frac{P}{\sigma^2}\right), \tag{1}$$

其中 P 是发送符号的功率。

中央服务器接收到的符号 $\hat{s}_i$ 首先通过重塑层调整为一个四维语义特征矩阵，记为 $\hat{k}_i \in \mathbb{R}^{4\times5\times22\times22}$，这一过程对应于信道解码。随后，采用由三个卷积核大小为 (3,3,3) 的三维卷积层组成的 3D-CNN，以及

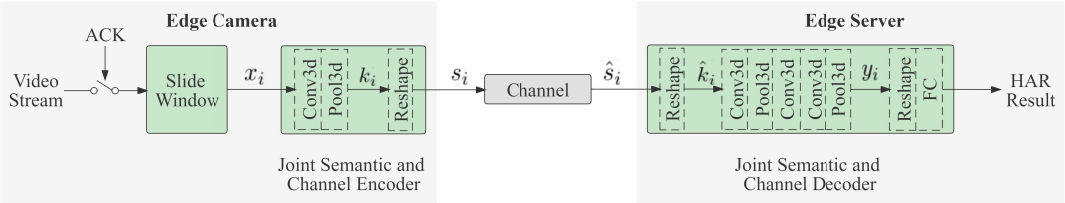

图3. 摄像头与服务器之间的收发器设计。

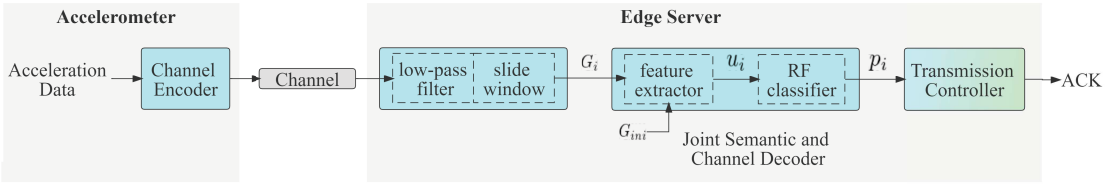

图4. 加速度计与服务器之间的收发器设计。

填充为 (1, 1, 1) 的设置和两个卷积核大小为 (2,2,2)、步幅为 (2, 2, 2) 的三维最大池化层 [15], , 以提取更深层的语义特征 $y_i \in \mathbb{R}^{8\times1\times5\times5}$。我们将人体活动识别分类视为目标任务，其目的是识别 5 种可辨认的日常活动，包括：

- 睡觉：用户躺在卧室的床上。
- 休息：用户坐在客厅的沙发上看书或看电视。
- 穿衣：用户在卧室的衣柜前穿衣或脱衣。
- 进食：用户坐在桌旁或在厨房用餐。
- 打电话：用户坐在客厅的书桌旁打电话。

为获得此 HAR 分类结果，先通过重塑层将高维语义特征调整为一个向量，再经由全连接线性层输出所有 5 种活动的概率。概率最高的活动即为最终的 HAR 分类结果。

B. 加速度数据收发器

虽然利用视频数据能够在具备所有必要信息的情况下，准确预测 HAR 任务结果，但持续传输视频数据的成本很高。请注意，每种人体活动都包含一系列姿势转换，因此姿势转换意味着人体活动可能发生变化。我们利用用户佩戴的加速度计采集的加速度数据来检测姿势转换。检测到姿势转换时，服务器上的传输控制器会向相应的摄像头发送肯定确认信号，以启动监控视频的传输。为支持对上述 5 种日常活动进行分类，我们选择识别 4 种人体日常姿势，即躺卧、坐姿、站立和行走。

图4说明了用于传输加速度数据的收发器设计。加速度计采集3-轴向加速度模拟信号，模数转换器将这些信号转换为数字信号。由于加速度计的计算能力有限，这些数字信号未经任何进一步预处理，便由信道编码器直接编码用于无线传输。

在接收端通过信道解码器恢复后，三轴加速度信号输入截止频率为 0.3 Hz 的低通滤波器，以获得低频加速度信号，并通过近似重力加速度 [16] 来表征用户运动方向。随后，使用 1 秒滑动窗口对连续的三轴低频加速度信号进行采样，得到重力加速度矩阵，记为 G_i，其中 $G_i \in \mathbb{R}^{3\times50}$，其中 i 为时间序列号。令 $g_{ij} =\in \mathbb{R}^3$ 表示 G_i 中第 j 组三轴数值。我们假设 $g_{def} = [0, 0, 1] \in \mathbb{R}^3$ 为表示默认运动方向的默认重力向量。

然后，特征提取器根据 g_{def} 与 G_i 中每个向量 g_{ij} 之间的夹角，计算第 i 个余弦值 u_i，以表征运动方向变化程度，计算公式如下：

$$u_i = \frac{1}{50}\sum_{j=1}^{50}\frac{g_{ij}\cdot g_{def}}{|g_{ij}|\cdot|g_{def}|}. \tag{2}$$

不同的姿势可以通过运动方向变化程度加以区分。我们使用基于随机森林（RF）的分类器生成姿势分类结果 p_i，输入为 u_i 。

然后，我们将 p_i 输入传输控制器，通过比较当前姿势分类结果 p_i和之前的结果 p_{i-1}来检测姿势转换。为避免误检测，只有在检测到姿势转换后，姿势 p_i 持续超过 2 秒，该姿势转换才有效。姿势转换有效时，传输控制器会向

相应的边缘摄像头发送肯定 {v5}信号，作为上传视频语义特征的指令。

IV. 仿真

A. 设置

我们采用一个日常活动数据集进行仿真。该数据集中的视频、音频和数字数据（包括加速度和温度）由安装在智能家居环境中或佩戴在受监测人员身上的传感器同步采集 [17]。我们从该数据集中选取视频和三轴加速度数据这两种模态用于仿真。视频分别由安装在卧室、客厅和厨房的三台摄像头录制，三轴加速度数据则由佩戴在用户左腋下的加速度计采集。

在采集训练数据时，我们从原始视频中提取日常活动的视频片段，并获取相应的加速度数据。我们分别对视频片段和加速度原始数据应用所提出的视频采样和姿态特征提取方法，生成 14,199 个视频样本和 1,242 个姿态样本。对于人体姿态分类任务，姿态样本的标签包括“躺卧”“坐姿”“站立”和“行走”。对于 HAR任务，视频样本被标记为五种对应于智能家居环境中人类日常活动的类型，即“睡觉”“休息”“穿衣”“进食”和“打电话”，如前所述。所有标签均采用独热编码格式编码。我们按 7:3 的比例将样本划分为训练集和测试集。

我们使用 PyTorch 库实现所提出的方法。针对监控视频的语义收发器使用交叉熵损失函数在视频数据集上进行训练。初始学习率设为 0.003，每经过 4 个 epoch 就除以 4。批量大小为 32，训练在 15 个 epoch 后停止。姿态分类使用随机森林模型，并在加速度数据集上进行优化。

B. 性能

1）通信开销：我们比较了三种视频压缩方法的通信开销，即所提出的面向任务的语义通信框架（HAR-SC）、所提出的结合传输控制的 HAR-SC（HAR-SC-TC）以及 MPEG-4。

在 HAR-SC 中，通信开销 C 定义为完成目标任务所需的传输符号数量。HAR-SC 的通信开销 C_{SC} 通过以下方式计算：

$$C_{SC} = L \times N_f$$

其中， L 是传输符号向量的长度， N_f 是完成目标任务所需的符号向量总数。

在 HAR-SC-TC 中，只有检测到有效的姿势转换（PT）时才会上传视频语义特征，通信开销 C_{TC} 由下式计算：

$$C_{TC} = L \times N_t$$

表 I 不同压缩方法下的通信开销

方法	通信开销
MPEG-4	356,573,829（信噪比=7dB） 210,237,977（信噪比=13dB） 145,780,221（信噪比=19dB） 111,048,888（信噪比=25dB）
HAR-SC	8,963,680
HAR-SC-TC	174,240

其中， N_t 是目标任务所需的特征上传次数。

此外，我们假设信道带宽为 1Hz，并将 MPEG-4, C_{MPEG-4} 的通信开销表示为

$$C_{MPEG} = \frac{N_b}{\log_2\left(1 + 10^{\gamma/10}\right)}$$

其中， N_b 是由 MEPG-4 编码器压缩的视频数据大小，单位为比特； γ 是加性高斯白噪声信道的信噪比，单位为 dB。

我们使用了一个大小为 110MB、包含 29,640 帧的视频，并通过所提出的视频采样方法从中提取了 1,852 个语义特征。计算结果如表 I 所示。当信道信噪比分别为 7dB、13dB、19dB 和 25dB 时，与 MPEG-4 方案相比，HAR-SC 的通信开销分别降低了 97.5%、95.7%、93.8% 和 91.9%。与 HAR-SC 相比，HAR-SC-TC 的通信开销降低了 98.0%。这证明，所提出的任务导向通信框架能够显著降低通信开销。

2）分类准确率：随后，我们根据活动分类准确率评估 HAR-SC 和 HAR-SC-TC 在 HAR 任务上的性能。对于 HAR-SC，基于 3D-CNN 的语义收发器在不同信道条件下进行训练，其中信噪比根据公式（1）计算。优化后的语义收发器分别在信噪比为 7dB、13dB、19dB 和 25dB 的条件下进行测试，结果如图 5 所示。请注意，当 SNR_{test} 远低于 SNR_{train} 时，准确率会下降，因为当信道噪声过大且训练时未遇到这种情况时，收发器无法纠正传输错误。我们还注意到，当 SNR_{test} 接近或高于 SNR_{train} 时，准确率趋于达到一个由模型训练决定的饱和值。仿真结果还表明，在较差信道条件下优化的收发器在信道条件不佳时性能更好，并且在信道条件变化时更加稳定。

对于 HAR-SC-TC，我们在所考虑的 HAR 任务中，以分类准确率为指标，测试了基于 3D-CNN 的语义收发器与用于加速度数据的基于射频的收发器相结合的方案。结果如表 II 所示，其中数值表示在房间内检测到某项活动的次数，后续括号中的数值表示分类准确率。
HAR-SC-TC 的上传频率远低于 HAR-SC，这意味着

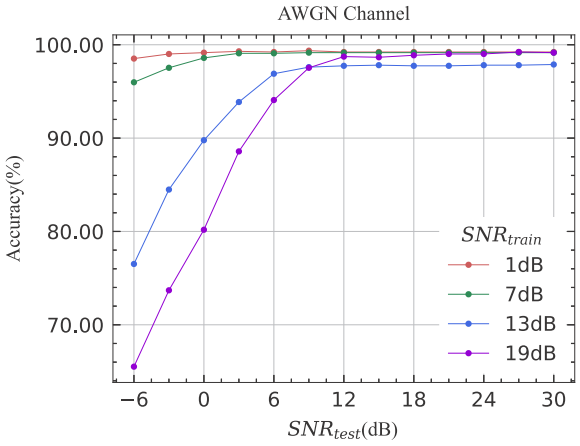

图 5. HAR-SC 在加性高斯白噪声信道上的分类准确率。

表 II HAR-SC-TC 的分类结果

活动 \ room	客厅	卧室	厨房
睡眠	0	10(100%)	0
rest	0	12(100%)	0
穿衣	0	2(100%)	0
eat	0	0	1(100%)
call	13（100%）	0	0

通信开销和功耗显著降低。此外，HAR-SC-TC 的分类准确率与 HAR-SC 一样高，这证明传输控制器模块在不降低目标任务性能的情况下减少了通信开销。需要说明的是，HAR-SC-TC 中用于监控视频的语义收发器和用于加速度数据的收发器是分别设计和优化的。这些结果验证了使用根据加速度数据预测的 PT 结果作为传输控制信号来决定视频信号传输的有效性。

五、结论

本文提出了一种面向任务的通信方法，用于将多个终端设备的多模态数据传输至中央服务器。具体而言，我们提出使用一种模态数据的接收信息来控制另一种数据速率高得多的模态数据的传输。我们以智能家居环境中的HAR任务为例。更具体地说，我们设计了一种基于 3D-CNN 的收发器，用于监控视频的语义通信并输出 HAR 结果；还设计了一种基于射频的收发器，用于传输加速度数据并输出 PT结果。PT结果用作传输控制信号，以决定是否将监控视频传输至服务器。数值结果表明，所提出的方法在不降低目标任务性能的情况下，相较于传统视频编码器实现了低得多的通信成本，从而验证了所提出的传输控制方法的有效性。

参考文献

[1] W. Weaver，“对通信数学理论的近期贡献”，ETC: a review of general semantics，第261–281页，1953年。[2] J. Bao、P. Basu、M. Dean、C. Partridge、A. Swami、W. Leland和J. A. Hendler，“迈向语义通信理论”，载于2011 IEEE Network Science Workshop，2011年，第110–117页。[3] C. E. Shannon，“通信数学理论”，The Bell system technical journal，第27卷，第3期，第379–423页，1948年。[4] N. Farsad、M. Rao和A. Goldsmith，“用于文本联合信源-信道编码的深度学习”，载于2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)，2018年，第2326–2330页。[5] Z. Weng和Z. Qin，“用于语音传输的语义通信系统”，IEEE Journal on Selected Areas in Communications，第39卷，第8期，第2434–2444页，2021年。[6] E. Bourtsoulatze、D. Burth Kurka和D. G¨und¨uz，“用于无线图像传输的深度联合信源-信道编码”，IEEE Transactions on Cognitive Communications and Networking，第5卷，579，2019 3，567第期。– 。[7] T. Han、Q. Yang、Z. Shi、S. He和Z. Zhang，“用于高效语音传输的语义保持通信系统”，IEEE Journal on Selected Areas in Co mmunications，第41卷，第1期，第245–259页，2023年。[8] Z. Zhang、Q. Yang、S. He、M. Sun和J. Chen，“借助多层次语义信息的图像无线传输”，载于2022 International Symposium on Wireless Communication Systems (ISWCS)，2022年，第1–6页。[9] X. Kang、B. Song、J. Guo、Z. Qin和F. R. Yu，“用于无人航空系统场景分类的面向任务的图像传输”，IEEE Transactions on Communications，第70卷，第8期，第5181–5192页，2022年。[10] Y. Zhang、W. Xu、H. Gao和F. Wang，“用于协作式目标识别的多用户语义通信”，载于2022 IEEE International Conference on Communications Workshops (ICC Workshops)，2022年，第157–162页。[11] H. Xie、Z. Qin、X. Tao和K. B. Letaief，“面向任务的多用户语义通信”，IEEE Journal on Selected Areas in Co mmunications，第40卷，第9期，第2584–2597页，2022年。[12] H. Xie、Z. Qin和G. Y. Li，“用于视觉问答的面向任务多用户语义通信”，IEEE Wireless Communications Letters，第11卷，第3期，第553–557页，2022年。[13] P. Jiang、C.-K. Wen、S. Jin和G. Y. Li，“用于具有HARQ的句子语义传输的深度信源-信道编码”，IEEE Transactions on Communications，2022年。[14] Jiang, Peiwen、Wen, Chao-Kai、Jin, Shi和Li, Geoffrey Ye，“用于视频会议的无线语义通信”，IEEE Journal o n Selected Areas in Communications，第41卷，第1期，第230–244页，2023年。[15] D. Tran、L. Bourdev、R. Fergus、L. Torresani和M. Paluri，“利用3D卷积网络学习时空特征”，载于Proceedings of the IEEE international conference on computer vision，2015年，第4489–4497页。[16] D. Curone、G. M. Bertolotti、A. Cristiani、E. L. Secco和G. Magenes，“一种基于三轴加速度计信号的实时自校准算法，用于检测人体姿势与活动”，IEEE Transactions on Information Technology in Biomedicine，第14卷，第4期，第1098–1105页，2010年。[17] A. Fleury、M. Vacher和N. Noury，“健康智能家居中日常生活活动的基于SVM的多模态分类：传感器、算法及初步实验结果”，IEEE Transactions on Information Technology in Biomedicine，第14卷，第2期，第274–283页，2010年。

该图“fig1.png”可从以下位置获取，格式为“PNG”：

http://arxiv.org/ps/2302.02608v1

此图“fig2.png”提供“PNG”格式版本，获取地址：

http://arxiv.org/ps/2302.02608v1

此图“fig3.png”提供“PNG”格式版本，获取地址：

http://arxiv.org/ps/2302.02608v1

此图“fig4.png”提供“PNG”格式版本，下载地址：

http://arxiv.org/ps/2302.02608v1