基于多模态融合与注意力机制的实时视频通信延迟补偿方法

李元好[1,2] 张如青[1,2] 薛 峰[1,2]
LI Yuanhao ZHANG Ruqing XUE Feng

摘 要 在实时视频通信中，网络环境往往复杂多变，导致视频数据的传输出现不同程度延迟，影响用户体验。为此，提出基于多模态融合与注意力机制的实时视频通信延迟补偿方法。利用多模态融合技术与注意力机制深度提取并强化了实时视频通信延迟中关键的延迟特征，以此准确预测并计算实时视频通信的延迟。基于这些精准数据，建立延迟补偿器模型，进一步优化数据帧处理路径，实现实时视频通信的延迟补偿。仿真测试表明，所提出的方法能够将视频通信整体延迟时间控制在 0.3 s 内，证明基于多模态融合与注意力机制的实时视频通信延迟补偿方法能够显著减少延迟，确保通信流畅无阻，给用户带来即时互动的极致体验。

关键词 多模态融合；注意力机制；实时视频；通信延迟；延迟补偿

doi：10.3969/j.issn.1672-9528.2024.11.023

0 引言

实时视频通信作为一种基于 IP 技术实现音视频实时交互的通信技术，允许多用户同时在互联网上进行低延迟、高清流畅的语音和视频通话[1]。然而，实时视频通信过程中常常面临网络延迟的问题，这不仅影响用户体验，还可能导致音视频同步失真、交互延迟增加等严重后果。因此，研究有效的实时视频通信延迟补偿方法具有重要意义。文献 [2] 提出了一种基于高速低载波比的实时视频通信延迟补偿方法，通过减少信号处理的复杂度和传输过程中的冗余数据，有效缩短视频数据的传输时间，从而降低端到端的延迟。不过，该方法需要较高的技术水平和复杂的算法支持，增加了系统的开发和维护成本；文献 [3] 提出了一种基于宽带相控阵的实时视频通信延迟补偿方法，该方法通过电子方式精确控制波束的方向和形状，减少信号干扰和提高通信质量，从而满足实时视频通信对带宽和延迟的严格要求。但由于宽带相控阵系统包含大量的阵列元素和信号处理单元，因此其功耗通常较大。在长时间运行的情况下，可能会对系统的散热和电源供应提出更高的要求。文献 [4] 则基于 5G 通信技术提出了一种实时视频通信延迟补偿方法。5G 网络采用了先进的资源调度算法，能够更有效地利用网络资源，降低网络拥塞和延迟。这对于完善实时视频通信中的延迟补偿至关重要。然而并非所有设备都支持 5G 网络，可能会遇到设备兼容性问题，这要求用户必须升级或更换设备以支持 5G 网络。为了解决上述成本高、功耗大和交通性差的问题，本文提出了一种基于多模态融合与注意力机制的实时视频通信延迟补偿方法，通过多模态融合与注意力机制，综合考虑视频、音频、文本等多种模态的数据信息，更全面地捕捉和理解视频通信中的关键信息，从而更准确地预测和补偿延迟，提升用户体验。

1 基于多模态融合与注意力机制的实时视频通信延迟补偿方法设计

1.1 基于多模态融合提取实时视频通讯延迟特征

为了从实时视频通信的视频帧内容、音频流特性、网络状态参数等多模态数据源中融合提取与通信延迟相关的特征，通过专用的数据采集软件将视讯软件近期的视频数据、音频数据、网络状态数据、文本数据以及传感器数据进行整理记录，并整合成一个数据集合。鉴于每种模态均承载着其独特的格式与表达逻辑，需要通过滤波等技术去除数据中的噪音[5]，并将预处理后的数据输入到各自模态专属的编码器中。各自模态专属的编码器会独立地抽取各实时视频通信延迟模态输入数据的深层特征。将这些单模

1. 郑州经贸学院大数据与人工智能学院 河南郑州 451191
2. 河南省多模态感知与智能交互技术工程研究中心
河南郑州 451191

[基金项目] 2023 年郑州经贸学院校级科研平台（多模态数据感知与融合技术工程研究中心）；2023 年度郑州经贸学院青年科研基金项目“基于二维码技术的高校多媒体教室智能管理系统的研究与实现”

态的特征映射至一个统一的语义子空间内，根据图 1 方法对特征数据进行融合。

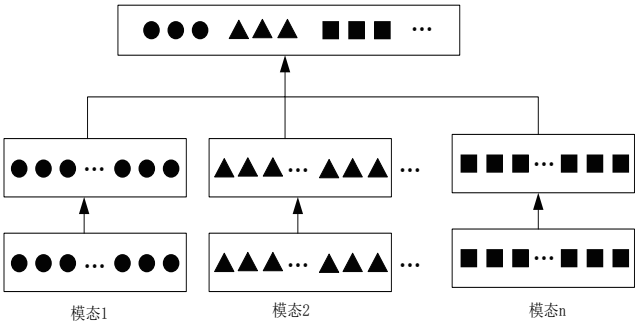

图 1 单模态的特征映射融合方法

假设图 1 中不同模态的特征映射向量分别为 $\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_n$，使用多模态融合公式将其进行融合，以此将提取的视频、音频和文本特征进行拼接，形成一个高维的特征向量，融合过程的计算公式为：

$$\boldsymbol{B} = c(\boldsymbol{a}_1 m_1 + \boldsymbol{a}_2 m_2 + \ldots + \boldsymbol{a}_n m_n) \tag{1}$$

式中：$\boldsymbol{B}$ 代表融合后的特征向量，c 代表将不同模态的特征映射向量转换为共享语义子空间的参数，m 代表不同模态的权重系数，n 代表映射特征数量。

1.2 基于注意力机制增强关键实时视频延迟特征

在提取出的多模态实时视频通信延迟特征中，并非所有特征都对延迟预测具有同等重要性 [6]。因此，本文利用注意力机制，动态地学习和分配不同特征的重要性权重 m。沿着不同的映射空间方向，将实时视频延迟特征分解为多个不同的特征编码，同时利用平均池化操作分别捕获特征映射方向，具体计算公式为：

$$\begin{cases} d_k^1 = \dfrac{1}{h_1}\sum b_1 \\ d_k^2 = \dfrac{1}{h_2}\sum b_2 \\ \cdots \\ d_k^n = \dfrac{1}{h_n}\sum b_n \end{cases} \tag{2}$$

式中：d_k^n 代表在第 k 个通道的输出特征映射方向，h 代表此方向的映射高度。根据输出的特征映射方向，全连接层会生成一个预测分布结果，将其引入注意力权重生成模块中，模块的全局最大池化层会整合成一个空间信息，生成紧凑的向量集。此时 Sigmoid 激活函数会将紧凑向量集转换为注意力权重向量，并确保权重值在 0 到 1 之间，以此动态地调整不同特征映射的贡献度。使用逐元素乘法将每个特征映射中的每个元素与其对应的注意力权重相乘，此时具有高注意力权重的特征元素将被放大，从而在后续处理中占据更大的比重；相反，具有低注意力权重的特征元素将被抑制，其影响在后续层中将被减弱，实现关键实时视频延迟特征的增强，并对非关键特征进行抑制。

1.3 预测计算实时视频通信延迟

基于增强后的关键实时视频延迟特征，对实时视频通信延迟差值进行预测计算，以此指导后续的延迟补偿操作，确保视频通信的流畅性和实时性。将处理后的特征数据集划分为训练集、验证集和测试集，并使用训练集数据训练神经网络模型。对模型中每个神经元的输入与权重相乘并加上偏差，得到神经元的输出。将当前层的输出作为下一层的输入，通过前向传播计算得到预测值。然而在此过程中可能会出现过拟合情况 [7]，需要使用验证集监控训练过程，并采取早停、正则化、Dropout 等措施，解决过拟合的问题。

将得到的预测时刻与当前时刻的延迟进行计算，得到这两个时刻之间的延迟差值。这个差值代表了从当前状态到预测状态的传输延迟变化量，具体的计算公式为：

$$\Delta F = \left\| F(g) - F(g+1) \right\| \tag{3}$$

式中：ΔF 代表实时视频通信延迟，$F(g)$ 代表在 g 时刻的通讯网络传输延迟数值，$F(g+1)$ 则代表 $g+1$ 时刻的通信网络传输延迟数值。

1.4 建立实时视频通信延迟补偿器模型

直接应用实时视频通信延迟，建立实时视频通信延迟补偿器模型，以此补偿部分网络延迟和其他因素引起的视频通信延迟。将计算得到的实时视频通信延迟数值输入 PID 控制中，动态调整通信网络的控制参数，并通过比例子控制器、积分子控制器以及微分子控制器 3 个子控制器计算通信网络控制量 [8]，具体计算公式为：

$$\begin{cases} H_1 = \mathrm{d}\alpha\Delta F \\ H_2 = \dfrac{1}{g}\int \Delta F z \mathrm{d}g \\ H_3 = g\dfrac{z\Delta F}{zg} \end{cases} \tag{4}$$

$$L = H_1 + H_2 + H_3 \tag{5}$$

式中：L 代表在时刻 g 的延迟补偿器控制量，H_1、H_2 和 H_3 分别代表该时刻 3 个子控制器的输出值，α 代表微分时间常数，z 则代表积分时间常数。将控制量输入延迟补偿器模型中，补偿器模型会根据这些控制量及其内在的逻辑和算法，将这些控制量转化为具体的网络传输参数调整指令，以减少或抵消实际传输过程中的延迟，延迟补偿器模型输出公式为：

$$K=\beta(iL+\int \Delta F \mathrm{d}\chi) \tag{6}$$

式中：K 代表延迟补偿器模型输出结果，β 代表实时视频通信信号延迟补偿系数，i 代表实时视频通信信号的衰减因子，χ 代表数据传输的通信长度。延迟补偿器模型会将输出结果制定成相应的补偿策略，动态调整视频流的编码速率、发送间隔等传输参数，以减小延迟对实时视频通信的影响。

1.5 优化实时视频通信数据帧处理路径

在审视实时视频通信延迟现象时，只利用实时视频通信延迟补偿器模型无法解决全部问题[9-10]。为此，本文还对实时视频通信数据帧处理路径进行了优化，解决了通信链路节点在单位时间周期内通信时隙与视频数据流之间的精准协同。对视频数据源到视频接收端的数据帧传输最短路径进行规划，规划结果如图 2 所示。

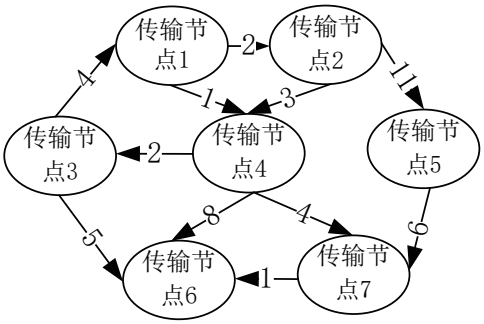

图 2 数据帧传输规划路径

从未被访问的节点中挑选出距离当前节点最近的邻接传输节点。每当一个新的传输节点被选定为传输路径的一部分时，立即更新该节点到起始点的最短距离记录，并标记该节点为已访问，不断重复上述过程，以此更新和优化整个传输路径。

确定了最优传输路径后，对实时视频数据进行高效封装，形成适合网络传输的数据帧，封装格式如表 1 所示。

表 1 视频数据封装格式

视频文件格式	视频封装格式
.avi	AVI（Audio Video Interleaved）
.wmv、.asf	WMV（Windows Media Video）
.mpg、.mpeg、.vob、.dat、.3gp、.mp4	MPEG（Moving Picture Experts Group）
.mkv	Matroska
.rm、.rmvb	Real Video
.mov	QuickTime File Format
.flv	Flash Video

这些数据帧将沿着最优路径，通过多传输节点快速、准确地传递至目标接收端，完成实时视频通信数据帧处理路径的优化，结合实时视频通信延迟补偿器模型，实现实时视频通信延迟补偿方法的设计。

2 仿真测试

2.1 测试准备

将 MM-IMDB-WIKI 影视短剧多分类数据集作为样本，输入搭建好的 Wireshark 的 VoIP 和 RTP 分析插件中，并对其分辨率、帧率、码率以及延迟等数据进行设置，具体参数数值如表 2 所示。将测试平台安装在 CPU 为 Intel Core i7-10700K，8 核 16 线程，主频 3.8 GHz；内存为 32 GB DDR4 RAM，2666 MHz；网络带宽为上行 1 Gbit/s、下行 1 Gbit/s 的计算机上，并按照预设的参数启动仿真测试，模拟实时视频通信过程。使用本文基于多模态融合与注意力机制的实时视频通信延迟补偿方法、基于高速低载波比的实时视频通信延迟补偿方法和基于宽带相控阵的实时视频通信延迟补偿方法，对样本进行延迟补偿，观察并记录应用延迟补偿算法后的整体延迟时间。

表 2 实时视频通信延迟补偿方法测试样本参数

样本	分辨率	帧率	码率	发送延迟 /s	传播延迟 /s	处理延迟 /s
1	1920×1080	60	5000	0.52	6.54	8.49
2	1280×720	30	3000	2.98	3.25	5.16
3	3840×2160	30	10 000	7.33	9.76	3.66
4	3840×2160	60	4000	4.12	4.83	1.88
5	1920×1080	60	6000	1.83	1.78	4.62
6	1920×1080	60	8000	9.84	2.92	1.58
7	1280×720	30	5500	1.98	7.81	3.25
8	1920×1080	30	2500	1.34	4.78	5.98
9	1920×1080	60	12 000	5.38	9.97	3.32
10	1920×1080	60	3500	1.98	6.15	2.58
11	1920×1080	60	7000	7.16	8.65	8.49
12	3840×2160	60	9000	7.65	6.19	7.92
13	1280×720	30	6000	8.16	4.98	3.64
14	1280×720	30	3000	5.66	1.68	2.79
15	3840×2160	60	11 000	4.97	1.16	3.65

2.2 测试结果与分析

将不同的实时视频通信延迟补偿方法的测试结果进行整理，具体参数如图 3 所示。根据图 3 的结果可知，经过本文设计的基于多模态融合与注意力机制的实时视频通信延迟补

偿方法补偿后的视频通信整体延迟时间可以控制在 0.3 s 内。相比之下，基于高速低载波比的实时视频通信延迟补偿方法和基于宽带相控阵的实时视频通信延迟补偿方法虽然在一定程度上也能缓解延迟问题，但其效果却远远无法与本文方法相媲美。这表明本文方法在降低延迟方面表现出色，远优于其他延迟补偿技术，展示了多模态融合与注意力机制在实时视频通信延迟补偿领域的高效的数据处理能力，能够应对大规模用户和复杂网络环境的需求，有效缓解网络抖动、丢包以及带宽限制等因素对实时视频通信质量的影响，保持通信的稳定性和流畅性。

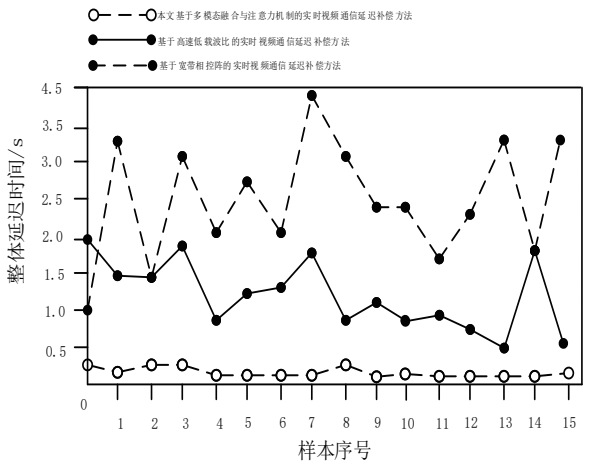

图 3 实时视频通信延迟补偿方法测试结果

3 结语

本文通过对基于多模态融合与注意力机制的实时视频通信延迟补偿方法的研究，为实时视频通信提供了更丰富的数据维度，解决了依赖单一的视频流或音频流进行信息传输的问题，优化了数据传输和处理流程，使实时视频通信能够在海量数据中自动学习和分配不同数据源，显著降低了实时视频通信过程中的延迟。尽管基于多模态融合与注意力机制的实时视频通信延迟补偿方法在提升数据维度、优化传输流程以及降低延迟方面展现出显著优势，但其仍存在一些不足。例如，该方法在实现多模态数据的融合与处理时，可能会面临较高的计算复杂度和资源消耗，对硬件设备的性能要求较高。此外，尽管注意力机制能够自动学习和分配不同数据源的重要性，但在某些特定场景下，其准确性可能受到噪声、网络波动等外界因素的干扰，从而影响延迟补偿的效果。因此，在实际应用中，仍需不断探索和优化该方法，以提升其稳定性和适用性。未来，随着技术的发展，需要在实时视频通信的延迟补偿中，不断引入跨领域的融合创新，满足日益增长的多元化需求，优化传输算法和策略，实现全局最优的延迟控制。

参考文献：

[1] 王一帆 , 张雪芳 . 基于多模态视频分类任务的模态融合策略研究 [J]. 计算机科学 , 2024, 51 (S1): 501-505.

[2] 何昊 , 郝振洋 , 俞强 . 高速低载波比下控制延迟分析与补偿策略研究 [J]. 微特电机 , 2023, 51 (8): 9-15.

[3] 张祺 . 宽带相控阵延时补偿系统设计方法 [J]. 电子信息对抗技术 , 2024, 39 (1): 52-55.

[4] 韦举敏 . 基于 5G 通信技术的延迟容忍网络数据传输方法 [J]. 信息与电脑 (理论版), 2023, 35 (6): 213-215.

[5] 张虎成 , 李雷孝 , 刘东江 . 多模态数据融合研究综述 [J]. 计算机科学与探索 ,2024,18(10):2501-2520.

[6] 王铁君 , 张泽宇 , 郭晓然 , 等 .MCFA-UNet：结合多尺度融合与注意力机制的图像生成网络 [J/OL]. 计算机工程与应用 ,1-14[2024-04-23].http://kns.cnki.net/kcms/detail/11.2127.tp.20240710.2144.009.html.

[7] 曾水玲 , 李昭贤 , 张嘉雄 , 等 . 结合注意力机制和编码器—解码器架构的化学结构识别方法 [J]. 中国图象图形学报 , 2024, 29 (7): 1960-1969.

[8] 黄鸿殿 , 李倩 , 黄原 , 等 . 基于因子图的 AUV 协同导航通信延迟误差补偿算法 [J]. 无人系统技术 , 2024, 7 (1): 115-123.

[9] 杨成林 , 周勋 , 严新荣 . 采用流式并行架构处理数据帧的优化方法 [J]. 舰船电子工程 , 2024, 44 (1): 76-80.

[10] 吴红海 , 马华红 , 邢玲 , 等 . 一种多用户协作博弈的视频机会传输路由算法 [J]. 软件学报 ,2020,31(12):3937-3949.

【作者简介】

李元好（1992—），通信作者（email：365252909@qq.com），男，河南汝州人，硕士，助教，研究方向：计算机应用。

张如青（1993—），女，河南开封人，硕士，讲师，研究方向：计算机应用。

薛峰（1983—），男，河南郑州人，硕士，副教授，研究方向：算法、计算机应用。

（收稿日期：2024-07-28）