

内容自适应量化参数决策的视频感知编码方法

祁海洋, 晏星宇, 郑永康, 徐 龙

(宁波大学人工智能学院, 宁波 315211)

摘 要:以第三代音视频编码标准(Audio video coding standard 3, AVS3)为代表的现代视频编码标准显著提升了压缩效率,但传统优化范式仍高度依赖均方误差(Mean squared error, MSE),而MSE与人眼视觉系统(Human visual system, HVS)的感知特性相关性较弱。近年来,基于知识蒸馏的方法通过生成感知引导的量化参数(Quantization parameter, QP)映射,有效改善了压缩视频的视觉质量。然而,现有方法大多局限于全帧内编码配置,难以直接推广至同时包含帧内与帧间预测的实际视频编码场景。针对上述问题,本文提出一种轻量级的内容自适应QP决策方法。在编码流程之前部署视频内容预分析模块,利用预训练卷积神经网络(Convolutional neural network, CNN)提取输入视频的显著时空特征;随后,决策模型依据这些特征动态确定不同视频片段所适用的编码策略。实验结果表明,所提方法能够针对不同类型的视频内容选择更有效的编码策略。与AVS3基准编码器相比,该方法在多尺度结构相似性(Multi-scale structural similarity, MS-SSIM)评价指标下平均降低7.11%的Bjontegaard Delta码率(Bjontegaard Delta rate, BD-rate)。进一步将其集成至高度优化的AVS3内容感知编码器后,在基本保持峰值信噪比(Peak signal-to-noise ratio, PSNR)性能的同时,MS-SSIM BD-rate相对于共同的AVS3高性能模型(High performance model, HPM)12.0基准进一步改善4.79个百分点。需要指出的是,当前方法采用片段级二值门控,尚不支持帧/块粒度精细调控。

关键词: 视频编码;第三代音视频编码标准;人眼视觉系统;深度学习;知识蒸馏

中图分类号: TN919.81

文献标志码: A

引用格式: 祁海洋, 晏星宇, 郑永康, 等. 内容自适应量化参数决策的视频感知编码方法[J]. 数据采集与处理, 2026, 41(5):1306-1317. QI Haiyang, YAN Xingyu, ZHENG Yongkang, et al. Content-adaptive quantization parameter decision method for perceptual video coding[J]. Journal of Data Acquisition and Processing, 2026, 41(5):1306-1317.

引 言

近年来,4K/8K超高清视频、虚拟现实和增强现实(Augmented reality, AR)等高保真视觉数据快速增长,对现代视频压缩技术提出了前所未有的挑战。以我国第三代音视频编码标准(Audio video coding standard 3, AVS3)^[1]和国际H.266多功能视频编码(Versatile video coding, VVC)标准^[2]为代表的新一代视频编码标准,仍沿用基于块的混合编码框架^[3-4]。该框架通过帧内/帧间预测、变换、量化和熵编码等环节,逐步消除视频中的空间冗余与时间冗余。其中,率失真优化是编码器的核心机制,其目标是在给定码率约束下最小化重建失真。

然而,传统率失真优化(Rate-distortion optimization, RDO)通常采用均方误差(Mean squared error, MSE)或峰值信噪比(Peak signal-to-noise ratio, PSNR)作为失真度量。尽管MSE具有数学形式简单、

计算开销低等优点,但其难以准确刻画人眼视觉系统(Human visual system, HVS)的非线性感知特性^[5-7]。因此,针对MSE进行优化的编码器往往会将过多码率分配给视觉上不敏感的纹理区域,同时可能导致视觉上敏感区域的块效应、振铃和模糊等失真。

为缓解上述客观失真度量与主观视觉感知之间的不一致,感知视频编码已成为重要研究方向^[8]。在多种技术路线中,基于知识蒸馏的方法^[9-12]表现出较大潜力。此类方法通常由感知优化的“教师”神经网络生成帧级或块级量化参数(Quantization parameter, QP)映射,将其码率分配知识迁移至传统“学生”编码器,从而在保持标准解码器兼容性的前提下改善主观视觉质量。

现有知识蒸馏框架主要面向图像数据集或全帧内编码的视频帧序列。若将此类静态方法直接应用于包含复杂时间预测关系的视频序列,往往难以获得稳定收益。实验观察表明,不同视频在纹理密度、运动规律和局部亮度等方面差异显著,单一且固定的感知优化策略难以适用于所有内容。对于高频纹理密集或运动随机性较强的序列,过于激进的感知QP调整甚至可能导致客观编码性能明显下降。因此,感知优化需要具备内容感知能力,即根据局部视频特性动态启用或调整感知模块^[13-14]。

针对上述问题,本文提出一种面向AVS3视频编码的内容自适应QP决策方法,并将轻量级预分析模块无缝嵌入编码流程。该模块通过分析输入视频的视觉特征,自适应决定是否启用基于知识蒸馏的感知优化,从而在提升感知编码效率的同时增强整体性能的稳定性。

本文的主要贡献如下:

(1)提出一种内容自适应QP决策框架,显著增强了知识蒸馏感知优化方法在视频编码场景中的稳定性与鲁棒性。

(2)设计一种高效的两阶段感知优化控制流程,首先采用预训练的16层视觉几何组网络(Visual geometry group 16-layer network, VGG-16)提取视频内容特征,随后利用轻量级逻辑回归模型完成低开销的启用/关闭决策。

(3)分别在AVS3参考软件和高度优化的内容感知编码器(Content-aware codec, CAC)上验证所提框架,在多种视频内容上获得了稳定且可叠加的编码收益。

1 相关工作

在深度学习广泛应用之前,研究者已开展了大量将感知特性融入传统混合编码框架的工作。感兴趣区域(Region of interest, ROI)编码通过将更多码率分配给人脸等关键语义区域,提高重要内容的重建质量。与此同时,基于HVS的方法依据局部纹理复杂度、亮度方向性等空间掩蔽效应调整局部QP。VVC标准中引入的感知量化参数调整(Perceptual QP adjustment, PerceptQPA)工具即通过测量高频交流系数能量估计块级纹理复杂度^[15-16]。此类方法虽能在一定程度上改善主观质量,但通常依赖人工设计特征和经验规则,因而优化能力受到限制。

近年来,深度学习从两条主要技术路线推动了视频压缩的发展。第一条路线是端到端学习式视频压缩。以变分超先验结构^[17]为基础,此类方法使用神经网络直接将像素映射为紧凑码流,并可结合多尺度结构相似性(Multi-scale structural similarity, MS-SSIM)^[18]、学习感知图像块相似度(Learned perceptual image patch similarity, LPIPS)^[6]等感知损失进行联合优化。然而,完全神经网络化的编码架构与现有硬件生态兼容性有限,工程部署成本较高。第二条路线是深度学习辅助的传统混合视频编码。该类方法不改变标准的基本编码框架,而是利用神经网络增强其中的特定模块,例如预处理去噪、基于CNN的环路滤波以及重建后质量增强网络(如增强型超分辨率生成对抗网络(Enhanced super-resolution generative adversarial network, ESRGAN)^[19])。本文研究的知识蒸馏方法属于这一范畴^[11-12]:将深层感知信息转化为符合标准的QP映射,从而获得显著压缩收益。然而,如何在帧内与帧间联合编码场景中稳定应用该方法,仍是尚未充分解决的问题,也是本文的主要研究动机。

进一步地,近期端到端神经视频编码通过多样化时空上下文、特征调制以及实时化设计持续提升压缩效率与工程实用性^[20-22],相关研究也开始面向无人机等实际应用场景验证端到端智能视频压缩的可行性^[23]。这些进展表明,学习式编码在时域建模、码率控制和实际部署方面不断成熟,但其与传统标准编码器的协同及面向具体内容的轻量级决策仍具有研究空间。

2 内容自适应感知编码决策方法

2.1 现有方法的局限性

知识蒸馏感知优化方法(如文献[11])在实际部署中的主要障碍是其难以适应高度动态的视频内容。当编码器面对快速运动或复杂高频纹理场景时,静态、无条件启用的优化策略容易造成不合理的码率分配,进而引发明显的率失真性能下降。因此,有必要引入自适应的内容感知门控机制。

为直观说明这一问题,图1给出了感知优化有效与无效场景的代表性示例。图中第1行为感知QP调整结果的灰度可视化,其中灰度越黑表示对应区域的QP降低幅度越大,即编码器为该区域分配更多码率;灰度越白则表示QP增加幅度越大,即相应减少码率分配;第2行为对应的视频内容,图中标注的百分数表示启用感知优化后相对于AVS3基准编码器的MS-SSIM Bjøntegaard Delta码率(Bjøntegaard Delta rate, BD-rate)^[24]变化,负值表示获得感知编码收益,正值表示编码性能出现退化。在有效场景中(图1(a)),感知驱动的QP调整能够较好地匹配视频内容特性,从而获得明显的感知编码收益;而在无效场景中(图1(b)),相同的QP调整策略由于与局部时空特性不匹配,反而可能造成率失真性能下降。上述现象表明,固定启用的感知优化策略难以保证不同视频内容下编码性能的稳定性,因此有必要引入内容自适应决策机制。

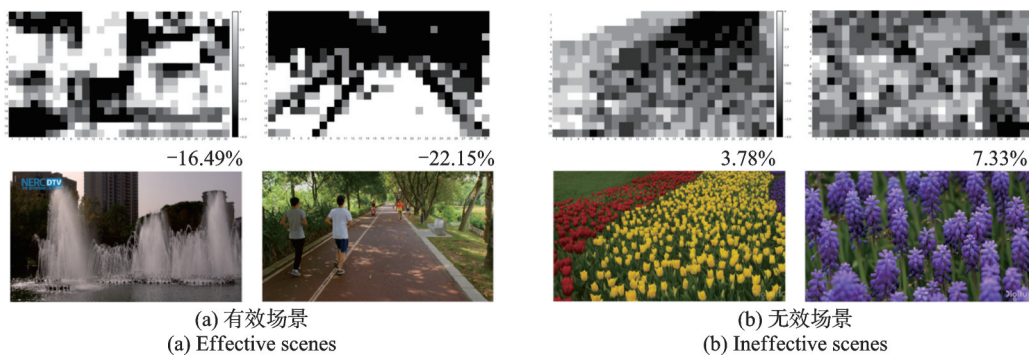


图1 不同场景下的感知QP调整及编码性能对比

Fig.1 Perceptual QP adjustment and coding performance comparison under different scenes

2.2 框架设计

为克服静态启用策略的不足,本文在AVS3编码器前端部署内容自适应决策框架,并将其作为预分析模块,如图2所示。在进入实际块级编码循环之前,该模块扫描视频序列的前若干帧,提取感知描述特征,并输出二值控制信号,以动态决定是否启用后续的感知优化模块。

下游感知优化模块沿用文献[11]中的知识蒸馏框架。该框架采用两个预训练端到端图像编码模型,分别以MS-SSIM和MSE为优化目标。首先估计并归一化两个模型的码率分布图,再通过二者的比值获得反映局部感知重要性的映射,并结合缩放因子 β 将其进一步转换为块级QP映射,从而引导AVS3编码器将更多比特分配至人眼视觉较为敏感的区域。本文保留该感知QP生成方式,而未直接将其中的端到端图像编码模型替换为端到端视频编码模型,主要是因为本文关注的是已有感知QP优化

方法从全帧内编码扩展至帧内/帧间联合编码后出现的内容依赖性和性能不稳定问题,其核心目标是研究“何时启用”感知优化,而非重新设计感知QP生成模型。保留原有模块有利于控制实验变量,更直接地评价所提内容自适应决策机制的作用。此外,端到端视频编码模型通常涉及运动估计、参考帧以及跨帧潜变量等时域预测机制,其内部码率分配与AVS3块级QP控制之间仍需建立额外的映射关系。因此,本文采用模块化设计,将内容自适应决策与感知QP生成过程相互解耦,以降低系统复杂度并保持较好的编码器兼容性。

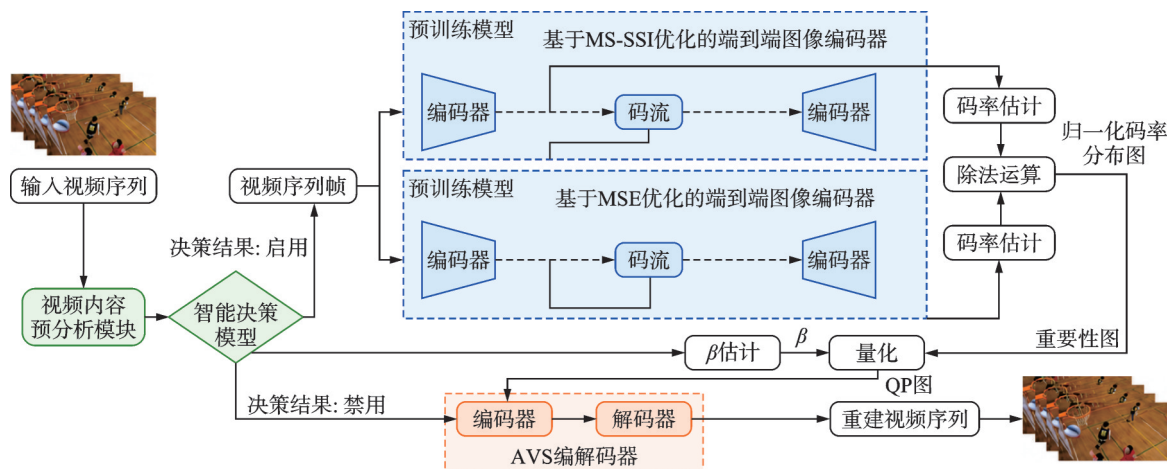


图2 内容自适应决策的整体架构

Fig.2 Overall architecture of the content-adaptive decision

2.3 视频内容预分析

视频内容预分析模块的目标是在实际编码开始之前,以较低计算开销获得能够反映视频时空特性的紧凑描述,并将其作为后续智能决策模型的输入。考虑到单一网络层难以同时描述纹理细节、局部结构和高层语义,本文采用基于深层卷积表示的多尺度特征提取方式,从不同网络深度获得互补信息。

已有研究表明,预训练CNN特征具有较好的通用视觉表征能力^[25]。如图3所示,本文采用预训练VGG-16^[26]作为骨干网络,并仅分析每个视频片段的前 $N=10$ 帧,以控制预分析开销。对于第 t 帧,分别从浅层、中层和深层提取特征 f_{low} 、 f_{mid} 和 f_{high} 。其中,浅层特征主要保留边缘、颜色和纹理等细粒度空间信息;中层特征用于描述局部结构模式;深层特征用于表征较高层次的视觉语义。随后,通过多尺度特

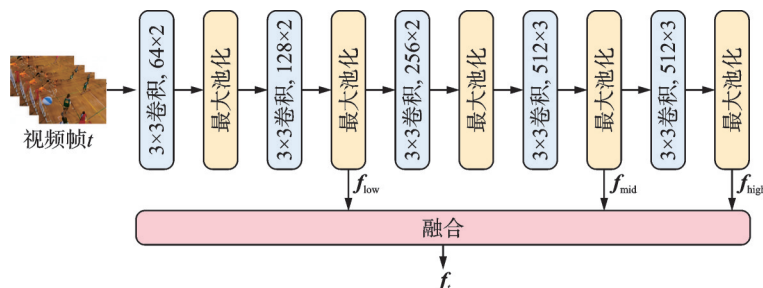


图3 基于VGG-16骨干网络的多尺度特征提取模块流程

Fig.3 Flowchart of the multi-scale feature extraction module based on the VGG-16 backbone network

征融合以及相邻帧特征变化统计,将原始深层表示进一步归纳为纹理、运动、复杂度和运动一致性4个视频内容指标。

由于不同网络层的通道维度不一致,首先分别对3个层级的特征图执行全局平均池化(Global average pooling, GAP),再将其拼接,得到第 t 帧的标准化高维多尺度表示向量

$$\mathbf{f}_t = [\mathbf{f}_{\text{low}}, \mathbf{f}_{\text{mid}}, \mathbf{f}_{\text{high}}] \quad (1)$$

为将抽象深层特征转化为可解释的视频内容特征,本文进一步通过维度归一化与时序池化,将原始向量序列 $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_N\}$ 压缩为4个核心感知指标,即纹理、运动、复杂度和运动一致性。

纹理指标用于表征视频空间细节的总体响应强度。对于第 t 帧,本文根据多尺度特征向量各维响应的平均值构造帧级纹理指标,并通过时间维度上的平均得到序列级纹理特征,即

$$\text{Texture}_t = \frac{1}{D} \sum_{i=1}^D |\mathbf{f}_{t,i}|, \quad \text{Texture} = \frac{1}{N} \sum_{t=1}^N \text{Texture}_t \quad (2)$$

式中 D 为特征向量维度。

运动指标用于刻画相邻帧语义特征的变化速度,其逐帧归一化位移与序列级均方根平均为

$$\text{Motion}_t = \|\mathbf{f}_{t+1} - \mathbf{f}_t\|, \quad \text{Motion} = \sqrt{\frac{1}{N-1} \sum_{t=1}^{N-1} \text{Motion}_t^2} \quad (3)$$

复杂度用于衡量特征内部的离散程度,定义为帧内标准差的均值,即

$$\text{Complexity} = \frac{1}{N} \sum_{t=1}^N \text{std}(\mathbf{f}_t) \quad (4)$$

运动一致性用于衡量时间维度上的稳定程度,通过下式有界化

$$\text{Consistency} = \frac{1}{1 + \text{Motion}} \quad (5)$$

经上述处理,视频序列最终被压缩为四维感知特征向量

$$\mathbf{v} = [\text{Texture}, \text{Motion}, \text{Complexity}, \text{Consistency}] \quad (6)$$

2.4 智能决策模型

智能决策模型接收视频内容预分析模块输出的四维特征向量,并根据当前视频片段的内容特性判断是否启用后续感知优化模块。具体而言,定义映射 $F: \mathbf{v} \rightarrow y$,其中 $y \in \{0, 1\}$ 。当 $y = 1$ 时启用感知优化模块,并利用生成的块级QP图指导AVS3编码;当 $y = 0$ 时不启用感知优化模块,视频按照原始AVS3编码路径进行处理。为获得离线训练所需的可靠标签,本文使用标准高效视频编码(High efficiency video coding, HEVC)测试序列(A~F类,覆盖不同分辨率和时空特性)构建多样化训练集。具体地,将每个序列划分为互不重叠的32帧片段,共得到292个训练样本。对每个片段 S_i ,在相同编码配置下分别进行两次编码:一次启用感知优化,另一次使用标准AVS3基准编码器。采用MS-SSIM评价指标下的BD-rate衡量率失真性能,并按式(7)生成真值标签 y_i 。

$$y_i = \begin{cases} 1 & \text{BD-rate} \leq 0 \\ 0 & \text{其他} \end{cases} \quad (7)$$

为降低在线决策的计算开销,本文采用轻量级逻辑回归(Logistic regression, LR)分类器。给定标准化特征向量 $\mathbf{v}_i$,启用概率为

$$p_i = \sigma(\mathbf{w}^T \mathbf{v}_i + b) \quad (8)$$

式中 $\sigma(\cdot)$ 为标准Sigmoid函数, $\mathbf{w}$ 和 b 分别为权重向量与偏置。当 $p_i \geq 0.5$ 时触发二值决策 $\hat{y}_i = 1$ 。模型参数通过最小化二元交叉熵损失与L2正则项 $\|\mathbf{w}\|^2$ 之和进行学习,以抑制过拟合,即

$$L=-\frac{1}{M}\sum_{i=1}^M[y_i\log p_i+(1-y_i)\log(1-p_i)]+\lambda_r\|\boldsymbol{w}\|^2 \tag{9}$$

式中: M 为样本数, λ_r 为正则化系数。经五折分层交叉验证,所提LR决策模型的准确率为92.3%,精确率为82.0%,召回率为93.3%, F_1 值为87.3%。其中,较高的召回率表明模型对能够从感知编码中获益的序列具有很高的敏感性。这一设置符合实际应用需求,即宁可少量误启用,也不应漏掉可获得显著感知增益的内容。

3 实验结果与分析

3.1 实验设置

所提框架集成于AVS3高性能模型(High performance model,HPM)12.0参考软件。全部实验均采用随机接入(Random access,RA)配置。为评估典型分发码率下的感知编码性能,基础QP设置为{38,45,52,60}。决策模型训练集采用不同分辨率的标准HEVC测试序列(A~F类),其内容覆盖多种纹理复杂度、运动特性和空间分辨率,有利于构建具有一定内容多样性的训练样本;HEVC测试序列仅作为视频内容来源,所有训练样本的标签仍通过AVS3 HPM 12.0在启用和禁用感知优化两种配置下分别编码并比较MS-SSIM BD-rate获得。性能测试采用AVS3通用测试条件(Common test conditions,CTC)中的代表性4K超高清(Ultra high definition,UHD)序列。采用不同来源的标准测试序列进行训练和测试,可以避免相同视频内容同时出现在训练集与测试集中,并在一定程度上考察决策模型对未知视频内容的泛化能力。所有AVS3测试序列在训练阶段均未出现。此外,为检验框架应对快速场景切换的能力,将若干不同的标准视频片段按时间顺序拼接,构造了一个64帧4K合成序列。在所测试QP范围内,采用BD-rate评价率失真性能,并分别使用客观PSNR指标和感知MS-SSIM指标综合衡量重建质量。

3.2 实验一:特征消融分析

为分析不同内容特征对决策性能的贡献,在保持训练数据、五折分层交叉验证方式以及逻辑回归模型参数不变的条件下,分别从完整四维特征向量中移除纹理、运动、复杂度和运动一致性特征,并重新训练决策模型。表1给出了不同特征组合下的分类性能。由表1可以看出,完整四维特征组合取得了最佳的综合分类性能。移除运动一致性特征后,模型性能下降最为明显,准确率和 F_1 分别由92.3%和87.3%下降至87.9%和81.4%,表明视频时间维度上的运动稳定性对于判断感知QP优化是否适用具有重要作用。移除运动特征后, F_1 下降至82.6%,进一步说明动态内容变化是影响感知优化有效性的重要因素。相比之下,纹理和复杂度特征的移除造成的性能下降相对较小,但完整特征组合仍取得最优结果。上述结果表明,4类特征能够从空间细节和时间变化等不同方面描述视频内容,并具有一定的互补性。

表1 不同特征组合下的决策模型分类性能

Table 1 Classification performance of the decision model with different feature combinations				%
特征组合	准确率	精确率	召回率	F_1
全部特征	92.3	82.0	93.3	87.3
去除纹理	90.2	79.1	91.5	84.8
去除运动	88.8	76.6	89.7	82.6
去除复杂度	91.4	80.5	92.2	86.0
去除运动一致性	87.9	75.2	88.6	81.4

3.3 实验二:标准 AVS3 基准上的性能

首先,将所提自适应框架(AVS3+Ours)与标准 AVS3 基准编码器 HPM 12.0 进行比较,并分别计算 PSNR 与 MS-SSIM 指标下的 BD-rate,以评价客观重建保真度和感知质量。

表 2 和图 4 给出了总体率失真性能,平均而言,AVS3+Ours 在 PSNR 指标下节省 3.39% 的 BD-rate,在 MS-SSIM 指标下节省 7.11% 的 BD-rate。对于 ParkRunning3 和 Fountains 等纹理丰富、运动较强的序列,MS-SSIM BD-rate 分别降低 22.15% 和 16.49%。同时,对于 Tango2 等序列,所提方法仍保持了较好的客观编码性能,其 PSNR BD-rate 为-8.30%。

表 2 AVS3+Ours 相对 AVS3 HPM 12.0 基准的 BD-rate 性能

Table 2 BD-rate performance of AVS3+Ours relative to the AVS3 HPM 12.0 anchor			%
测试序列	BD-rate (PSNR)	BD-rate (MS-SSIM)	
ParkRunning3	3.77	-22.15	
Fountains	-0.92	-16.49	
Campfire	-0.72	-8.21	
FolkDance	-4.54	-4.65	
Tango2	-8.30	-0.18	
Bosphorus	-6.32	-2.65	
StampCarving1	-7.34	-0.25	
FlowerPan	-2.55	-2.33	
平均	-3.39	-7.11	

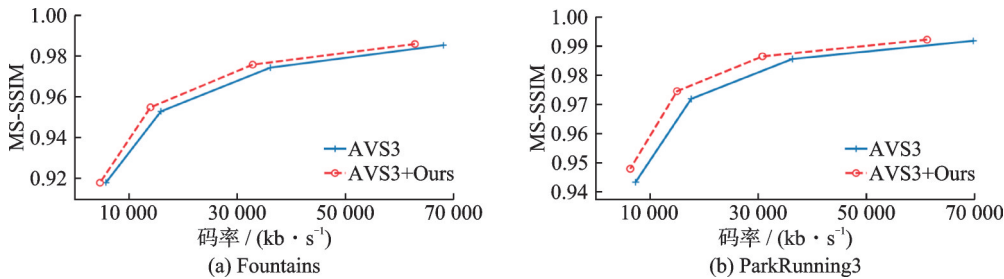


图 4 MS-SSIM 准则下的率失真性能对比

Fig.4 Rate-distortion performance comparison under the MS-SSIM criterion

3.4 实验三:优化 AVS3 编码器上的叠加增益

为验证所提框架与已有先进编码工具的兼容性及其在多工具联合配置下的可叠加性,本文进一步将所提方法集成至内容感知编码器(CAC)。CAC 是基于 AVS3 混合编码框架构建的高度优化编码平台,主要面向下一代 AVS 视频编码标准开展新型编码工具的技术预研与性能验证。该平台集成了智能自适应分辨率编码、基于模板匹配的帧内预测、虚拟参考帧预测、非局部结构环路滤波、轻量级卷积神经网络环路滤波以及深度超分辨率上采样等优化工具。在保持 CAC 原有编码工具配置不变的基础上,本文增加所提出的内容自适应 QP 决策模块(CAC+Ours),用于考察该模块能否在已有高性能编码平台上进一步获得编码增益。

如表 3 所示,尽管 CAC 基准已具备较强的压缩性能,集成所提框架后的 CAC+Ours 相对于 HPM

12.0 仍可进一步提升平均性能:PSNR BD-rate 由 -3.60% 改善至 -7.01% , MS-SSIM BD-rate 由 -21.45% 改善至 -26.24% 。对于 ParkRunning3 等视觉特性变化较为剧烈的序列,相对于共同 HPM 基准, CAC+Ours 的 MS-SSIM BD-rate 较 CAC 进一步改善 23.48 个百分点。对于感知调节可能无益的序列(如 ChefCooking 和 FlowerPan),决策模型能够关闭感知模块,从而保持 CAC 基准的编码效率并减少潜在性能退化。如图 5 所示, RD 曲线进一步验证了上述稳定增益。

表 3 CAC 与 CAC+Ours 相对 AVS3 HPM 12.0 的 BD-rate 对比
Table 3 BD-rate comparison of CAC and CAC+Ours relative to AVS3 HPM 12.0 %

测试序列	PSNR		MS-SSIM	
	CAC	CAC+Ours	CAC	CAC+Ours
Fountains	-1.11	2.66	-14.48	-24.25
ChefCooking	-33.68	-33.68	-39.31	-39.31
CityAlley	-0.13	-0.13	-12.26	-12.26
ParkRunning3	-10.83	-23.47	-11.93	-35.41
Campfire	-24.19	-25.11	-31.63	-39.19
FolkDance	-5.35	-9.89	-23.60	-25.30
FlowerPan	20.45	20.45	-6.01	-6.01
StampCarving1	13.25	10.25	-28.59	-28.95
Tango2	9.16	-4.14	-25.22	-25.47
平均	-3.60	-7.01	-21.45	-26.24

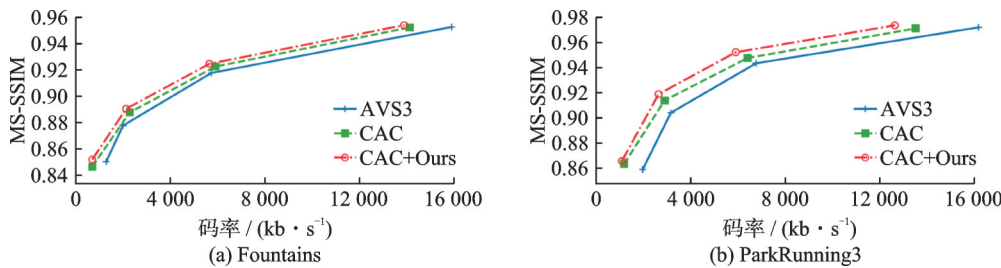


图 5 CAC 与 CAC+Ours 在 MS-SSIM 准则下的率失真性能

Fig.5 Rate-distortion performance of CAC and CAC+Ours under the MS-SSIM criterion

3.5 实验四:场景切换条件下的性能

为初步考察框架在突发场景切换条件下的行为,本文在合成拼接序列上比较了 3 种编码配置:标准 AVS3 基准、无条件感知蒸馏(Unconditional perceptual distillation, UPD)以及本文所提出的内容自适应框架。其中,UPD 作为静态基准,始终启用文献[11]中的感知优化模块。

如表 4 和图 6 所示,静态 UPD 策略在 MS-SSIM 指标下可节省 6.79% 的 BD-rate,但在 PSNR 指标下增加了 1.71% 的 BD-rate,说明固定码率分配策略在场景切换条件下存在客观质量风险。相比之下,在该合成序列上,所提内容自适应框架实现了 11.82% 的 MS-SSIM BD-rate 节省,同时避免了客观质量下降,其 PSNR BD-rate 相对基准略有降低(-0.53%)。上述初步结果表明,自适应切换机制有助于缓解静态感知调节在时间不连续场景中引起的性能波动。

表4 场景切换序列上的率失真性能对比

方法	QP	码率/($\text{kb}\cdot\text{s}^{-1}$)	PSNR/dB	MS-SSIM	BD-rate (MS-SSIM)/%	BD-rate (PSNR)/%
AVS3	38	15 257.80	39.72	0.983 1	—	—
	45	8 004.67	37.90	0.971 8		
	52	4 076.14	35.66	0.953 9		
	60	2 581.64	33.30	0.927 3		
UPD	38	17 523.53	40.02	0.982 9	−6.79	1.71
	45	7 119.53	37.52	0.972 2		
	52	3 740.59	35.21	0.953 0		
	60	2 541.40	33.60	0.928 7		
Ours	38	15 415.26	39.74	0.984 2	−11.82	−0.53
	45	7 421.18	37.54	0.973 3		
	52	3 410.28	35.25	0.952 7		
	60	2 504.64	33.31	0.927 0		

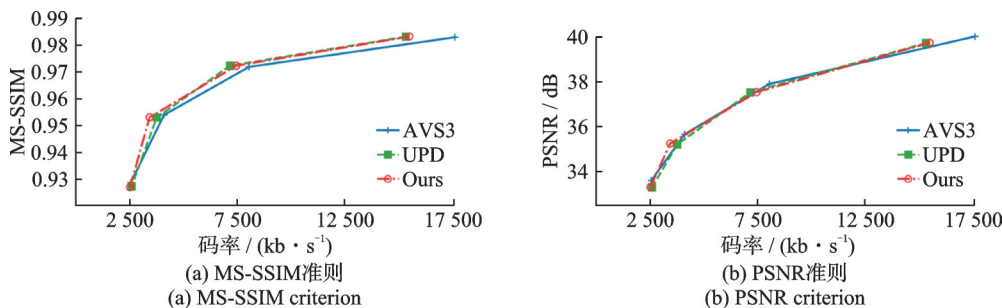


图6 AVS3、UPD与本文所提方法在合成场景切换序列上的率失真性能对比

Fig.6 Rate-distortion performance comparison of AVS3, UPD, and the proposed method on the synthetic scene-switching sequence

4 结束语

针对知识蒸馏感知优化方法在不同视频内容上性能不稳定的问题,本文提出了一种内容自适应决策框架。通过在预分析阶段集成多尺度特征提取器和轻量级逻辑回归模型,所提方法能够动态评估视频的时空特性,并据此决定是否启用感知优化模块。实验结果表明,所提方法具有较好的有效性与可扩展性。在标准 AVS3 基准编码器上,该框架能够同时改善平均感知指标和客观指标;集成至高度优化的 CAC 编码器后,仍可获得稳定的可叠加编码增益。对于包含突发时间切换的合成序列,初步实验也表明,自适应切换机制能够降低静态感知调节带来的率失真风险。与近期端到端神经视频编码工作^[20-23]不同,本文不替换传统混合编码框架,而是通过编码前的轻量级内容决策控制已有感知 QP 优化模块,更侧重于提升现有 AVS3 编码流程中的内容适应性与工程兼容性。需要指出的是,当前方法采用片段级二值决策,仅判断感知优化模块的启用或禁用,尚不能实现帧级或块级的精细调控,其在更丰富视频内容上的泛化能力仍有待进一步验证。后续可进一步探索以端到端视频编码模型作为教师网络,将时域感知信息更直接地引入感知 QP 生成过程。总之,本文所提框架在编码效率与内容适应性之间

取得了较好平衡,在不改变现有视频编码标准码流结构的情况下,可作为一种简单实用的编码器侧优化工具集成至现有混合视频编码框架中,以提升视频感知质量和压缩效率。

参考文献:

- [1] ZHANG J, JIA C, LEI M, et al. Recent development of AVS video coding standard: AVS3[C]//Proceedings of 2019 Picture Coding Symposium (PCS). Ningbo, China: IEEE, 2019: 1-5.
- [2] BROSS B, CHEN J, OHM J R, et al. Developments in international video coding standardization after AVC, with an overview of versatile video coding (VVC)[J]. Proceedings of the IEEE, 2021, 109(9): 1463-1493.
- [3] SULLIVAN G J, OHM J R, HAN W J, et al. Overview of the high efficiency video coding (HEVC) standard[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(12): 1649-1668.
- [4] BROSS B, WANG Y K, YE Y, et al. Overview of the versatile video coding (VVC) standard and its applications[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(10): 3736-3764.
- [5] MIN X, DUAN H, SUN W, et al. Perceptual video quality assessment: A survey[J]. Science China Information Sciences, 2024, 67: 211301.
- [6] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 586-595.
- [7] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: From error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.
- [8] ZHANG Y, ZHU L, JIANG G, et al. A survey on perceptually optimized video coding[J]. ACM Computing Surveys, 2023, 55(12): 1-37.
- [9] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. (2015-03-09). <https://arxiv.org/abs/1503.02531>.
- [10] GOU J, YU B, MAYBANK S J, et al. Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.
- [11] YANG R, LIU D, MA S, et al. Perceptual quality-oriented rate allocation via distillation from end-to-end image compression[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2024, 20(7): 1-22.
- [12] YANG R, LIU D, MA S, et al. Knowledge distillation from end-to-end image compression to VVC intra coding for perceptual quality enhancement[C]//Proceedings of 2021 IEEE International Conference on Image Processing (ICIP). Anchorage, AK, USA: IEEE, 2021: 3438-3442.
- [13] 何小海, 李鑫磊, 魏海涛, 等. 深度学习驱动的视频编码: 方法、进展与展望[J]. 数据采集与处理, 2026, 41(2): 515-542.
HE Xiaohai, LI Xinlei, WEI Haitao, et al. Deep learning driven video coding: Methods, progress, and perspectives[J]. Journal of Data Acquisition and Processing, 2026, 41(2): 515-542.
- [14] WU J, DENG D, LI Y, et al. Low-complexity content-aware encoding optimization of batch video[J]. Journal of Visual Communication and Image Representation, 2024, 104: 104295.
- [15] BOSSE S, HELMRICH C, SCHWARZ H, et al. Perceptually optimized QP adaptation and associated distortion measure: JVET-H0047[R]. Macao, China: Joint Video Exploration Team (JVET), 2017.
- [16] HELMRICH C, SCHWARZ H, MARPE D, et al. Improved perceptually optimized QP adaptation and associated distortion measure: JVET-K0206[R]. Ljubljana, Slovenia: Joint Video Exploration Team (JVET), 2018.
- [17] BALLÉ J, MINNEN D, SINGH S, et al. Variational image compression with a scale hyperprior[C]//Proceedings of the 6th International Conference on Learning Representations (ICLR). Vancouver, Canada: ICLR, 2018.
- [18] WANG Z, SIMONCELLI E P, BOVIK A C. Multiscale structural similarity for image quality assessment[C]//Proceedings of the 37th Asilomar Conference on Signals, Systems & Computers. Pacific Grove, CA, USA: IEEE, 2003, 2: 1398-1402.
- [19] WANG X, YU K, WU S, et al. ESRGAN: Enhanced super-resolution generative adversarial networks[C]//Proceedings of Computer Vision-ECCV 2018 Workshops. Munich, Germany: Springer, 2018: 63-79.

- [20] LI J, LI B, LU Y. Neural video compression with diverse contexts[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 22616-22626.
- [21] LI J, LI B, LU Y. Neural video compression with feature modulation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024: 26099-26108.
- [22] JIA Z, LI B, LI J, et al. Towards practical real-time neural video compression[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2025: 12543-12552.
- [23] 叶枫, 董凡可, 贾川民. 端到端智能视频压缩技术及其在无人机中的应用[J]. 数据采集与处理, 2025, 40(2): 303-319.
YE Feng, DONG Fanke, JIA Chuanmin. End-to-end video compression technology and its application in unmanned aerial vehicles[J]. *Journal of Data Acquisition and Processing*, 2025, 40(2): 303-319.
- [24] BJONTEGAARD G. Calculation of average PSNR differences between RD-curves: VCEG-M33[R]. Austin, TX, USA: ITU-T Video Coding Experts Group (VCEG), 2021.
- [25] SHARIF RAZAVIAN A, AZIZPOUR H, SULLIVAN J, et al. CNN features off-the-shelf: An astounding baseline for recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Columbus, OH, USA: IEEE, 2014: 806-813.
- [26] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of the 3rd International Conference on Learning Representations (ICLR). San Diego, CA, USA: ICLR, 2015.

作者简介:



祁海洋(2000-),男,硕士研究生,研究方向:深度学习、视频编码与智能编码, E-mail: 2411100165@nbu.edu.cn。



晏星宇(1992-),男,博士研究生,研究方向:视频编码、智能编码, E-mail: yanxingyu@nbu.edu.cn。



郑永康(2001-),男,硕士研究生,研究方向:深度学习、视频编码, E-mail: 2411100341@nbu.edu.cn。



徐龙(1977-),通信作者,男,博士,教授,研究方向:人工智能、计算机视觉、视频编码及人工智能与科学交叉研究, E-mail: xulong1@nbu.edu.cn。

(编辑:王静)

Content-Adaptive Quantization Parameter Decision Method for Perceptual Video Coding

QI Haiyang, YAN Xingyu, ZHENG Yongkang, XU Long*

(School of Artificial Intelligence, Ningbo University, Ningbo 315211, China)

Abstract: Modern video coding standards such as the audio video coding standard 3 (AVS3) have substantially improved compression efficiency, but conventional rate-distortion optimization still relies heavily on signal-fidelity measures such as mean squared error (MSE) and peak signal-to-noise ratio (PSNR), which do not fully reflect the nonlinear characteristics of the human visual system (HVS). Knowledge-distillation-based perceptual coding methods can improve visual quality by transferring bit-allocation knowledge from perceptually optimized learned image compression models to block-level quantization parameter (QP) maps. However, most existing approaches are designed and evaluated under image or all-intra coding configurations. When directly extended to practical video coding with both intra- and inter-frame prediction, a fixed perceptual QP adjustment strategy may produce content-dependent performance fluctuations, especially for sequences with complex textures or strong temporal variation. To address this problem, this paper proposes a lightweight content-adaptive QP decision method for AVS3 video coding. Before actual encoding, a video content pre-analysis module processes the first ten frames of each segment using a pretrained VGG-16 network. Features from shallow, intermediate, and deep layers are pooled and fused to describe fine-grained texture, local structure, and high-level semantics. Temporal variation of the fused features is further summarized into four compact descriptors: Texture, motion, complexity, and motion consistency. A logistic-regression classifier then generates a binary gating signal to determine whether the downstream perceptual optimization module is enabled or bypassed. The classifier is trained on 292 non-overlapping 32-frame segments constructed from standard HEVC test sequences, where the ground-truth label of each segment is determined by the MS-SSIM BD-rate difference between perceptual optimization and the AVS3 anchor under identical coding configurations. Five-fold stratified cross-validation yields an accuracy of 92.3%, a precision of 82.0%, a recall of 93.3%, and an F_1 score of 87.3%. Experiments on representative 4K UHD sequences under random-access configuration show that, relative to the AVS3 HPM 12.0 anchor, the proposed method reduces the average BD-rate by 7.11% under multi-scale structural similarity (MS-SSIM) and by 3.39% under PSNR. When integrated into a highly optimized content-aware code (CAC), the average MS-SSIM BD-rate relative to the common HPM anchor improves from -21.45% to -26.24% , corresponding to an additional gain of 4.79 percentage points, while the average PSNR BD-rate improves from -3.60% to -7.01% . On a synthetic 64-frame sequence containing abrupt scene changes, the proposed adaptive scheme achieves an 11.82% MS-SSIM BD-rate reduction and a -0.53% PSNR BD-rate, whereas unconditional perceptual distillation yields a 6.79% MS-SSIM reduction but a 1.71% PSNR increase. These results indicate that the proposed segment-level gating mechanism can retain perceptual coding gains while reducing the risk of objective-quality degradation caused by unconditional perceptual adjustment. The method does not alter the standard bitstream structure and can be integrated as a low-overhead encoder-side decision module in conventional hybrid video coding frameworks.

Highlights:

1. A content-adaptive gating framework is proposed for perceptual QP optimization in AVS3, enabling segment-level selection between perceptual optimization and the original coding path.
2. Multi-scale VGG-16 features extracted from only ten frames are summarized into four content descriptors, including texture, motion, complexity, and motion consistency, and a lightweight logistic-regression model performs the binary switching decision.
3. The proposed method achieves a 7.11% average MS-SSIM BD-rate reduction relative to HPM 12.0. When integrated into CAC, the MS-SSIM BD-rate relative to the common HPM anchor is further improved by 4.79 percentage points, while the scene-change experiment shows reduced objective-quality risk compared with unconditional perceptual optimization.

Key words: video coding; audio video coding standard 3 (AVS3); human visual system (HVS); deep learning; knowledge distillation

Foundation item: The "Pioneer" and "Leading Goose" R & D Program of Zhejiang Province (No.2025C01035).

Received: 2026-07-03; **Revised:** 2026-08-21

***Corresponding author, E-mail:** xulong1@nbu.edu.cn.