摘要

随着信息技术与多媒体通信技术的快速发展，视频已成为互联网中最重要的数据类型之一，并广泛应用于在线教育、远程医疗、视频会议、智能监控以及沉浸式交互等领域。与此同时，视频分辨率、帧率及动态范围等关键指标的持续提升，使视频数据规模呈现爆炸式增长趋势，对存储与传输系统带来了巨大压力。在此背景下，如何在保证视觉质量的前提下实现高效的视频表示与压缩，成为视频处理与通信领域的重要研究问题。

围绕视频数据的内在特性，本文系统分析了影响视频表示与视觉质量的关键因素，包括空间分辨率、时间分辨率及视觉感知机制。其中，空间分辨率决定图像细节表达能力，时间分辨率影响运动连续性与流畅性，而人类视觉系统对亮度、色度及频率信息的非均匀敏感性直接影响主观质量评价。在高分辨率、高帧率及高动态范围视频不断普及的背景下，这些因素共同推动视频数据量快速增长，并对压缩技术提出更高要求。在此基础上，本文从冗余特性的角度分析了视频压缩的核心原理。空间冗余来源于像素间的高度相关性，使局部区域具有较强可预测性；时间冗余源于相邻帧之间的内容相似性，使当前帧能够通过参考帧进行有效预测；感知冗余则来源于人类视觉系统的选择性敏感特性，使部分信息的压缩不会显著影响视觉质量。三类冗余的叠加使视频具有较高的压缩潜力。尽管传统基于块的混合编码框架通过层次划分、复杂预测模式及滤波等技术有效压缩了空间与时间冗余，但在大幅运动、非刚体形变及遮挡等复杂场景下，其基于线性位移假设的建模能力仍然有限，导致帧间相关性利用不足、预测残差增大及压缩效率下降。对于传统视频编码在复杂运动建模与参考信息表达方面的不足，本文在混合编码框架内开展研究。

首先，针对运动估计精度受限的问题，提出了一种面向运动信息高效利用的帧间预测优化。该方法分析了现有标准中解码端运动矢量细化（Decoder-side Motion Vector Refinement，DMVR）与双向光流优化（Bi-directional Optical Flow，BIO）结果未被复用的局限性，并提出将其优化后的高精度运动信息引入空间运动矢量候选集合，使其参与后续块的运动矢量预测，从而在初始阶段获得更优估计。在不增加码率开销的前提下，该方法实现了解码端信息的有效复用，提升了运动信息利用效率。

针对传统帧间预测中参考帧与当前帧在复杂运动场景下匹配程度不足、导致时间冗余利用效率受限的问题，本文对于虚拟参考帧（Virtual Reference Frame，VRF）生成方法进行研究。通过对前后参考帧进行高精度对齐与时空一致性建模，生成在结构与运动特性上更接近当前帧的虚拟参考帧，并将其引入参考帧列表参与预测过程，

从而降低预测误差与残差能量。基于此，本文提出了一种局部–全局时空上下文建模的VRF生成网络，通过融合分层卷积特征与基于状态空间建模的Mamba模块，实现多尺度冗余特性的统一建模。其中，局部特征用于捕获纹理与边缘信息，Mamba模块用于建模长程依赖关系，从而增强跨帧相关性。在运动建模方面，引入从粗到细的多级光流估计机制：粗级光流用于捕获大尺度运动趋势，细级光流对复杂边界与形变区域进行精细修正，从而减少运动建模误差带来的冗余残留，提高帧间预测精度。

综上所述，随着视频数据规模的增长与应用需求的提升，高效挖掘视频中的多维冗余信息已成为提升编码性能的关键。本文从空间、时间与感知特性出发，系统研究了冗余建模与利用方法，并从运动优化与参考信息增强两个层面提出改进策略。通过DMVR与BIO机制实现运动信息复用，并通过VRF方法增强跨帧相关性，两者协同提升了冗余利用效率。研究表明，强化时空冗余建模并引入学习驱动的参考生成机制，可有效突破传统编码框架的性能瓶颈，为高效视频压缩提供新的技术路径。

关键词：视频编码，帧间预测，深度学习，虚拟参考帧

Inter Prediction Optimization for Video Coding

Zhaoyu Li (Computer Application Technology)
Directed by SE. Yun Jiang and Prof. Siwei Ma

ABSTRACT

With the rapid development of information technology and multimedia communication technologies, video has become one of the most important data types on the Internet and has been widely applied in fields such as online education, telemedicine, video conferencing, intelligent surveillance, and immersive interaction. Meanwhile, the continuous improvement of key indicators such as video resolution, frame rate, and dynamic range has led to an explosive growth in video data volume, imposing significant pressure on storage and transmission systems. In this context, how to achieve efficient video representation and compression while maintaining visual quality has become an important research problem in the fields of video processing and communication.

Focusing on the intrinsic characteristics of video data, this paper systematically analyzes the key factors affecting video representation and visual quality, including spatial resolution, temporal resolution, and the human visual perception mechanism. Specifically, spatial resolution determines the capability of representing image details, temporal resolution affects motion continuity and smoothness, and the non-uniform sensitivity of the human visual system to luminance, chrominance, and frequency information directly influences subjective quality evaluation. With the increasing prevalence of high-resolution, high-frame-rate, and high-dynamic-range videos, these factors collectively drive the rapid growth of video data and impose higher requirements on compression techniques. On this basis, this paper analyzes the fundamental principles of video compression from the perspective of redundancy characteristics. Spatial redundancy arises from the high correlation among neighboring pixels, making local regions highly predictable; temporal redundancy originates from the similarity between adjacent frames, enabling the current frame to be effectively predicted from reference frames; perceptual redundancy stems from the selective sensitivity of the human visual system, allowing certain information to be compressed without significantly affecting visual quality. The combination of these three types of redundancy provides substantial potential for video com-

pression. Although the conventional block-based hybrid video coding framework effectively compresses spatial and temporal redundancy through techniques such as hierarchical partitioning, sophisticated prediction modes, and filtering, its modeling capability based on the linear displacement assumption remains limited in complex scenarios involving large motion, non-rigid deformation, and occlusion. This limitation leads to insufficient utilization of inter-frame correlation, increased prediction residuals, and reduced compression efficiency. To address the limitations of traditional video coding in complex motion modeling and reference information representation, this paper conducts research within the hybrid coding framework.

First, to address the limited accuracy of motion estimation, the optimization of frame interpolation for efficient utilization of motion information is proposed. This method analyzes the limitation that the refined results of Decoder-side Motion Vector Refinement (DMVR) and Bi-directional Optical Flow (BIO) in existing standards are not reused, and proposes to incorporate the optimized high-precision motion information into the spatial motion vector candidate list, allowing it to participate in subsequent motion vector prediction for coding blocks, thereby obtaining more accurate initial estimates. Without introducing additional bitrate overhead, the proposed method achieves effective reuse of decoder-side information and improves the utilization efficiency of motion information.

To address the limitation that the matching between reference frames and the current frame is insufficient in complex motion scenarios in traditional inter prediction, which restricts the utilization efficiency of temporal redundancy, this paper investigates a Virtual Reference Frame (VRF) generation method. By performing high-precision alignment and spatiotemporal consistency modeling on the forward and backward reference frames, a virtual reference frame that is closer to the current frame in both structural and motion characteristics is generated and introduced into the reference picture list to participate in the prediction process, thereby reducing prediction error and residual energy. Based on this, a VRF generation network with local–global spatiotemporal context modeling is proposed, which integrates hierarchical convolutional features and a Mamba-based state space modeling module to achieve unified modeling of multi-scale redundancy characteristics. Specifically, local features are used to capture texture and edge information, while the Mamba module models long-range dependencies, thereby enhancing inter-frame correlation. For motion modeling, a coarse-to-fine multi-stage optical flow estimation mechanism is introduced: the coarse-level optical flow captures large-scale motion trends, while the fine-level optical flow performs precise refinement for complex boundaries and deformation regions, thereby reducing redundancy caused

by motion modeling errors and improving inter-frame prediction accuracy.

In summary, with the growth of video data and increasing application demands, efficiently exploiting multi-dimensional redundancy in video has become crucial for improving coding performance. Starting from spatial, temporal, and perceptual characteristics, this paper systematically investigates redundancy modeling and utilization methods, and proposes improvements from two aspects: motion optimization and reference information enhancement. By enabling motion information reuse through DMVR and BIO mechanisms, and enhancing inter-frame correlation through the VRF method, the two approaches jointly improve redundancy utilization efficiency. Experimental results demonstrate that strengthening spatiotemporal redundancy modeling and introducing learning-driven reference generation mechanisms can effectively overcome the performance limitations of traditional coding frameworks, providing a new technical pathway for efficient video compression.

KEYWORDS: Video coding, inter prediction, deep learning, virtual reference frame