摘要

当今世界处于以信息化全面引领创新、以信息化为基础重构国家核心竞争力的新阶段，正迎来新一轮信息革命浪潮。其中产生的海量视频图像数据无法完全交由有限的人力分析，必须借助智能化机器视觉算法进行理解。然而现有面向人眼感知的视频编码技术受限于非语义层面的率失真建模，难以满足大数据时代下智慧城市和物联网等新型应用场景的机器视觉分析性能需求，因此面向机器视觉的图像视频编码研究应运而生。现有面向机器视觉的图像视频编码方法往往更关注于单一任务上的性能提升，对不同任务之间的相关性建模不足，难以满足复杂机器视觉应用场景的高效编码需求。因此，探索机器视觉任务之间的语义相关性，精准构建任务之间的预测模型，探索更加紧凑高效的任务特征表示，对于面向机器视觉的图像视频编码的编码效率提升有着重大意义。本文针对面向机器视觉的图像视频编码方法，从机器视觉任务之间的语义相关性建模出发，深入讨论了传统混合框架中任务相关性建模难题、多任务相关性的精细建模优化难题和单任务不同分布之间的语义相关性建模难题，最终提升了面向机器视觉任务的编码效率。本文主要创新点包括以下三个方面：

第一，提出了一种基于信号重要度的图像编码方法。针对传统混合编码框架中多任务联合建模困难的问题，提出了一种与多个机器视觉任务语义相关度高的代理任务并为其设计分析网络。基于该网络的梯度分析对输入图像每一个像素块在代理任务上的贡献进行数值建模，提出了机器视觉感兴趣区域和像素级机器视觉重要度两个概念。基于这两个机器视觉重要度的概念，设计了机器视觉引导的码率分配策略，实现信号驱动的非均匀码率分配。同时，设计基于深度学习的多上下文特征失真，结合率失真优化理论，构建机器视觉率失真模型并部署到传统混合编码框架中。与 H.266 的帧内编码相比，所提方法在图像分类、目标检测、语义分割和姿态估计四种机器视觉任务的 Top-1 准确率、平均精度均值上实现最高 29.19% 的性能增益。

第二，提出了一种多任务驱动的高效视频编码方法。针对多个机器视觉任务对应的特征联合精细建模优化困难的问题，受深度学习理论中的多任务学习启发，设计了由任务共享网络模块和任务特异性网络模块组成的多任务学习网络。以动作识别和视频重构两大任务为目标，对相应特征进行联合建模并构建多任务损失，通过端到端的学习策略分步进行训练优化。一方面，共享网络模块的存在充分利用了两个任务对应特征之间的语义相关性，降低了编码所需的码率消耗；另一方面，任务特异性网络模块分别提升了两个任务对应评价指标上的性能，同时实现了高保真度的视频重构与高准确度的动作识别分析。相比于 H.265 的低延时配置编码，本方法在实现与其相当的

MS-SSIM 性能的同时，在动作识别任务的 Top-1 准确率上实现了 106.99% 的性能增益。

第三，提出了一种文本提示词引导的高效图像编码方法。针对特征表示在语义层面不够紧凑和不同数据分布的机器视觉任务对应特征之间语义相关性建模困难两大问题，利用大规模预训练模型中蕴含的丰富知识信息，提出了文本模态特征表示和文本提示词引导的通用特征表示两种特征表示方法。其中，文本模态特征表示有效地提升了特征的语义紧凑度，进一步提升了编码效率。通用特征表示由大规模预训练模型提取，具有应对多种数据分布任务的能力，在此基础上通过文本提示词进行引导，实现语义信息的选择性表达，进而实现不同数据分布的机器视觉任务对应特征的联合建模。与 H.266 的帧内编码相比，所提方法在语义分割任务的平均精度均值上实现了 110.60% 的性能增益。

综上所述，本文提出了面向机器视觉的高效图像视频编码方法，从机器视觉任务之间的相关性建模出发，设计高效预测模型和优化策略，从信号数值分析角度和任务驱动优化角度实现了不同任务之间的精准预测建模。随后，利用大规模预训练模型中蕴含的丰富知识，对同一任务不同数据分布之间的语义相关性进行建模，进一步提升了面向机器视觉的图像视频编码的编码效率。本文从多个角度对机器视觉任务之间的相关性建模进行了探索研究，在多个机器视觉任务上显著提升编码性能，提升了面向机器视觉的图像视频编码在任务层面的通用性，为未来更加智能的面向机器视觉的图像视频编码研究奠定了基础。

关键词：视频编码，机器视觉，面向机器视觉的视频编码，特征编码，深度学习

High Efficiency Image/Video Coding for Machine

Zhimeng Huang (Computer Application Technology)
Directed by: Prof. Siwei Ma

ABSTRACT

Today's world has entered a new stage characterized by comprehensive innovation driven by informatization and the restructuring of national core competitiveness based on information technology, ushering in a new wave of the information revolution. The massive volume of produced image/video data cannot be fully analyzed by limited human resources and must rely on intelligent machine vision algorithms for efficient understanding. However, existing video coding technologies aimed at human perception are limited by non-semantic rate-distortion optimization and face challenges while meeting the requirements of machine vision analysis in new intelligent application scenarios in the era of big data, such as smart cities and the Internet of Things (IoT). Thus, research on image/video coding for machine has emerged as a necessity. Current image/video coding methods for machine vision focus more on the performance improvement for single tasks and lack sufficient modeling of correlations between different tasks, making it challenging to improve the coding efficiency of complex machine vision application scenarios. Therefore, exploring the semantic correlations between machine vision tasks, accurately constructing predictive models between tasks, and exploring more compact and efficient task feature representations are of great significance for improving the coding efficiency of image/video coding for machine. This dissertation proposes image/video coding methods for machine that starts with the modeling of semantic correlations between different tasks. It delves into modeling of correlations between different tasks in traditional hybrid coding frameworks, fine modeling optimization of multi-task correlations, and semantic correlation modeling between different distributions of a certain task to enhance the coding efficiency for machine vision. The main innovations of this dissertation include the following three aspects:

First, an importance-based image coding method is proposed. To address the difficulty of joint modeling of multiple tasks in traditional hybrid coding frameworks, a proxy task with high semantic relevance to several machine vision tasks is introduced, and an analysis network is designed for it. Gradient analysis based on this network models the contribution of each coding unit in the input image to the proxy task numerically, introducing the concepts of the Region of Interest for Machine (ROIM) and Pixel-Level Importance Score for machine (PLIS). Based on these concepts, a machine vision-guided bitrate allocation strategy is designed, achieving signal-driven non-uniform bitrate allocation. Simultaneously, a machine vision rate-distortion model is constructed by designing Multi-Context Feature Distortion (MCFD) based on deep learning, combined with rate-distortion optimization theory, and deployed in traditional hybrid coding frameworks. This method achieves a maximum performance gain of 29.19% in four machine vision coding tasks compared to the intra-frame coding of H.266.

Second, a multi-task driven efficient video coding method is proposed. Inspired by multi-task learning in deep learning theory, a multi-task learning network composed of task-shared modules and task-specific modules is designed. Targeting action recognition and video reconstruction, the features are jointly modeled, and a multi-task loss is constructed, optimized in stages through an end-to-end learning strategy. On the one hand, the existence of the shared module fully utilizes the semantic correlation between features corresponding to the two tasks, reducing the bitrate required for coding; on the other hand, task-specific modules improve the performance of the evaluation indicators corresponding to the two tasks, achieving high-fidelity video reconstruction and high-accuracy action recognition analysis. Compared to the low-delay configuration coding of H.265, this method achieves a performance gain of 106.99% in the top-1 accuracy rate of the action recognition task while achieving comparable MS-SSIM performance.

Third, a text-prompt guided efficient image coding method is proposed. Addressing the challenges of insufficiently compact semantic feature representations and the difficulty of modeling semantic correlations between features corresponding to different data distributions of machine vision tasks, this method leverages the rich knowledge contained in large-scale pretrained models. It proposes two types of feature representations: text-modality feature representation and text-prompt guided universal feature representation. The text-modality feature representation effectively improves the semantic compactness of features, further enhancing coding efficiency. The universal feature representation, extracted by large-scale pretrained models, has the capability to handle tasks with various data distributions. Guided by text prompts, it achieves selective expression of semantic information and joint modeling of features corresponding to different data distributions of machine vision tasks. This method achieves a performance gain of 110.60% in mean Average Precision(mAP) of semantic segmentation task compared to the intra-frame coding of H.266.

In summary, this dissertation proposes efficient image/video coding method for machine vision, starting from the modeling of correlations between machine vision tasks, designing efficient predictive models and optimization strategies, and implementing precise predictive modeling between different tasks from the perspectives of signal numerical analysis and task-driven optimization. Then, leveraging the rich knowledge contained in large-scale pretrained models, it models the semantic correlations between different data distributions of the same task, further enhancing the coding efficiency of image/video coding for machine vision. This dissertation explores the modeling of correlations between machine vision tasks from multiple angles, significantly improving performance in various machine vision tasks, enhancing the universality of image/video coding for machine at the task level, and laying a foundation for future more intelligent research in image/video coding for machine area.

KEY WORDS: Video coding, machine vision, video coding for machine, feature coding, deep learning