摘要

在以互联网和移动互联网为源头的第三次工业革命之后，人工智能技术正在推动第四次工业革命，为全球生产、生活带来变革。在多媒体产业领域中，数字化媒介作为信息载体，通过通信技术手段将世界相关联。随着可便携式拍照设备的普及和图像采集设备多维度质量提升，视觉信息数据量的快速增长，为其存储与传输带来空前挑战。在此背景下，基于深度学习的图像压缩方法由于其巨大压缩效率潜力以及其面向人眼视觉感知的优越适配性，成为当下学界和业界关注焦点。

过去，图像编码与机器视觉两个领域通常被独立研究，而随着深度学习成为两个领域共同的底层技术驱动以及面向智能分析的编码作为新兴需求的提出，学者们开始对两个领域研究目标的融合问题产生关注。本文从图像信息频域分层表征入手，探索基于分层形式的端到端图像编码技术与以压缩数据为输入的机器视觉分析表现，并讨论如何将信息论中率-失真函数与视觉分析任务联合优化分析。

本文主要工作可总结为如下三点：

- **提出基于频域分层的端到端图像压缩模型**。针对基于神经网络的图像编码模型缺乏可解释性问题，从信号分解角度提出了采取频域信号分解的端到端图像编码方法。通过多分辨率金字塔网络将原始图片信号分解到多频段维度上，并进一步在频域去除信息冗余。考虑到多频段信息融合问题，设计了采用非局部注意力机制的频域信息融合模块，实现对各频段信息自适应融合。在多数据集上，该模型实现主观质量和客观编码性能的提升，在 MS-SSIM 指标上超过下一代编码标准 H.266/VVC；同时，在基于重建图的机器视觉分析任务上，该编码方法展示出对语义相关信息的有效保留。
- **提出既考虑人眼视觉质量又兼顾下游机器视觉分析任务的分层式端到端图像压缩模型**。针对图像压缩与压缩域视觉分析相结合的问题，采用基于分层式生成的图像压缩方法并与下游视觉分析任务联合优化训练，得到压缩率、重建图像质量和下游分析任务性能最优权衡点对应模型。本文提出以压缩域数据为输入的多任务分析模型，进一步提升分析效率与减少资源消耗。本文方法在人脸数据集上进行了压缩与多种分析任务评估，实验结果展示本文方法在压缩率一定的情况下，基于压缩域数据的多任务分析方法可以取得与原始图像输入方法可比的分析结果，同时可以节省 99.6% 的传输码流（假设原图为 3 通道且比特位深度为 8）。
- **建模码率-失真函数与视觉分析任务间的联合关系，提出码率-失真-分类联合优**

化模型并对其统计特性进行分析。针对图像压缩与重建图视觉分析相结合的问题，提出码率-失真-分类联合优化模型（通过分类任务代表视觉分析任务）并对其统计特性进行论证。本文从特殊分布源上的特性开始，推导至一般分布源下的函数统计特性，即在一定条件下，码率-失真-分类具备单调性与凸函数性质。由此，本文通过在手写数字数据集上的实验说明在有损图像编码中，更低的码率会导致更高的像素级信息损失与重建图分类错误，实验结果符合码率-失真-分类的性质分析。

综上，本文从图像信号频域分解问题出发，探讨面向机器视觉的图像编码方法，提出既考虑人眼观感又考虑机器分析任务的分层式图像编码方法，进而说明以压缩域为输入的多任务分析网络与压缩模型的联合训练使得压缩数据在码率、失真、分析三个维度都取得了较好的效果。此外，针对面向视觉分析的图像编码方法提出码率-失真-分类联合优化建模并对其统计特性进行分析，为相关技术发展与理论讨论提供参考。

关键词：图像编码，机器视觉，信号分解，率-失真理论，深度神经网络

End-to-End Image Compression and Visual Analysis on Compressed Domain

Zhang Yuefeng (Computer Application Technology)
Directed by Prof. Ma Siwei

ABSTRACT

Following the third industrial revolution triggered by the Internet and mobile Internet, artificial intelligence technology driven by big data is driving the fourth industrial revolution, which will bring changes to global production and life activities. In the field of multimedia industry, digital media serves as information carrier to connect the world by the communication technology. With the popularity of portable cameras and multi-dimensional quality improvement of image collection devices, the rapid growth of visual information data volume brings unprecedented challenges for its storage and transmission. In this context, deep learning-based image compression methods have become the focus of academic and industrial fields due to their huge potential of compression efficiency and their superior adaptability for human visual system.

In the past, image coding and machine vision fields were usually studied separately, but with deep learning becoming the underlying technology driver for both fields and the emerging need for intelligent analytics-oriented coding, it has drawn scholars' attention to the convergence of research objectives in those two fields. In this thesis, we start from a hierarchical representation of image information in the frequency domain, explore end-to-end image coding techniques based on hierarchical forms with compressed data as input for machine vision analysis tasks, and discuss the joint optimization of the rate-distortion theory in information theory and the vision analysis tasks.

The main work of this thesis can be summarized as the following three points:

- **An end-to-end image compression model based on frequency transform is proposed.** To address the lack of interpretability of neural network-based image coding models, an end-to-end image coding method that adopts frequency-oriented transform is proposed from the perspective of signal decomposition. The original image signal is decomposed into multi-band dimensions by a multi-resolution pyramid network,

and the information redundancy is further removed in the frequency domain. Considering the multi-band information fusion problem, a frequency domain information fusion module using a non-local attention mechanism is designed to achieve adaptive fusion of information in each frequency band. The model achieves subjective quality and objective coding performance improvement on multiple datasets, exceeding the next-generation coding standard H.266/VVC on MS-SSIM metrics; meanwhile, the coding method demonstrates effective retention of semantically relevant information on reconstructed image-based machine vision analysis tasks.

- **A hierarchical end-to-end image compression model that takes into account both human eye vision quality and downstream machine vision analysis tasks is proposed.** To address the problem of combining image compression and vision analysis, a hierarchical generation-based image compression method is used and trained jointly with downstream vision analysis tasks to obtain a model corresponding to the optimal trade-off points of compression rate, reconstructed image quality and downstream analysis task performance. In this thesis, we propose a multi-task analysis model with compressed data as input to further improve the analysis efficiency and reduce resource consumption. The method is evaluated on a face dataset with both compression and multiple analysis tasks, and the experimental results demonstrate that the multi-task analysis method based on the compressed domain can achieve comparable analysis results with the original image input method under a certain compression rate, while saving 99.6% bitrates of the transmission stream (assuming the original image has 3 channels and its bit depth is 8).
- **The rate-distortion-classification joint optimization model is proposed and its statistical properties are analyzed by modeling the relationship between the rate-distortion and the visual analysis task.** For the problem of combining image compression with visual analysis on reconstructed images, the rate-distortion-classification joint optimization model (representing the visual analysis task by the classification task) is proposed and its statistical properties are justified. In this thesis, we start from the properties on special distribution sources and derive to the statistical properties on general distribution sources, i.e., under certain conditions, the rate-distortion-classification possesses monotonicity and convex function properties. Moreover, this thesis illustrates experimentally on handwritten digital datasets that in lossy image coding lower rates lead to higher pixel-level information loss and classification errors on the recon-

struction images, that experimental results are consistent with the statistical analysis of rate-distortion-classification.

In summary, this thesis discusses the image coding method for machine vision from the aspect of image signal decomposition, proposes a hierarchical image coding method that considers both human eye perception and machine analysis tasks, and then shows that the joint training of multi-task analysis network on compressed domain and compression model achieves better results in three dimensions: rate, distortion and analysis performance. In addition, the rate-distortion-classification joint optimization model is proposed for the visual analysis-oriented image coding method and its statistical characteristics are analyzed, providing a reference for related technical development and theoretical discussion.

KEY WORDS: Image Coding, Machine Vision, Signal Decomposition, Rate-distortion Theory, Deep Neural Network