摘要

人类社会正在迈向大数据和深度学习技术驱动的人工智能时代。图像和视频不再仅限于人类观看，而是更多地交给机器感知和理解其中的高层级语义信息，解决实际任务。由于传输带宽、存储空间受限，图像和视频要经过编码压缩以减少数据量，但编码引入的失真会改变机器的感知结果、降低任务性能。为解决这一问题，面向机器视觉的编码方法应运而生，相比面向信号保真或人类视觉感知优化的传统编码方法，在同等压缩率下能够获得更好的任务性能。然而，现有方法没有充分考虑多样的机器视觉模型对编码失真的感知特性及其差异，编码性能仍存在提升空间，欠缺适配不同机器视觉模型的通用性；也没有充分考虑人类和机器视觉系统感知特性的关联，难以兼顾人类和机器的统一感知需求。

本文将机器输出的视觉任务结果作为机器视觉感知的具象体现，从发现机器存在和人类相似的感知局限性出发，递进地提出了对图像信号中的机器视觉感知冗余、不同机器视觉模型的感知多样性、人类和机器视觉系统的感知相关性进行建模的方法。利用这些模型，针对不同内容的图像自适应地预测编码参数，缓解了现有方法性能、通用性、统一性受到限制的问题。本文的主要创新点包括：

第一，提出了机器视觉感知冗余的建模方法。研究了图像信号中机器视觉感知冗余的存在性及其特性，通过恰可识别失真（Just Recognizable Distortion，JRD）度量机器视觉模型感知结果不变或正确时所允许的最大编码失真，建立了机器视觉感知冗余和编码过程中图像级量化参数之间的联系，将这种冗余的建模问题转换为JRD预测问题。为了支撑对JRD的研究，构建了大规模的JRD自然图像数据集。考虑JRD的偏态分布特性，提出了基于集成学习的JRD预测框架，通过集成多个专家模型，在有参考和无参考图像的情况下准确地预测了JRD。基于预测的JRD自适应地决策编码参数，提升了面向机器视觉的编码性能。

第二，提出了机器视觉感知多样性的建模方法。研究了不同模型之间感知多样性的存在性及其对编码的影响，通过机器满意比（Satisfied Machine Ratio，SMR）表示编码前后感知结果一致的模型的占比，将这种多样性的建模问题转换为SMR预测问题。为了支撑对SMR的研究，建立了具有代表性的机器视觉模型库完成SMR标注，构建了大规模的SMR自然图像数据集。基于深度特征差异和SMR之间的非线性负相关性，提出了基于孪生神经网络的SMR预测模型，利用不同失真等级图像之间的SMR差异信息实现了数据增广学习，准确地预测了SMR。基于预测的SMR自适应地决策编码参数，提升了编码结果对于不同机器视觉模型的通用性。

第三，提出了人类和机器视觉感知相关性的建模方法。首先将这种相关性的建模问题转换为人类和机器满意比的统一预测问题。针对现有数据集规模无法支撑模型高效训练的问题，聚合多个具有代表性的图像质量评价模型模拟人类满意比的获取过程，生成后者的代理标签，实现了大规模数据集上的统一预训练。在预训练模型提取的多层级特征的基础上，设计了统一预测网络结构，通过差异特征残差学习、基于多头注意力的多层级特征聚合与池化、基于多层感知机混合结构的空间-通道信息融合模块，学习到了反映两种视觉系统感知相关性的紧凑特征，准确地预测了人类和机器满意比。基于预测的人类和机器满意比自适应地决策编码参数，提升了编码结果对于人类和机器感知需求的统一性。

关键词：面向机器视觉的视频编码, 感知编码, 恰可察觉失真, 恰可识别失真, 用户满意比, 机器满意比

Research on Distortion Perception Modeling for Machine Vision and Its Application in Coding

Qi Zhang (Computer Application Technology)
Supervised by Prof. Wen Gao and Prof. Siwei Ma

ABSTRACT

Human society is progressing toward an era of Artificial Intelligence driven by big data and deep learning technologies. Images and videos are no longer solely intended for human watching but increasingly utilized by machines to perceive and interpret high-level semantic information for solving practical tasks. Limited by transmission bandwidth and storage capacity, images and videos must be compressed to reduce data volume, which introduces distortions that alter machine perception and degrade task performance. To address this issue, Video Coding for Machines (VCM) has emerged. Compared to traditional methods optimized for signal fidelity or human visual perception, VCM methods achieve superior task performance under the same compression ratio. However, existing methods lack full consideration on the perceptual characteristics of diverse machines to coding distortion and their differences. Therefore, the coding performance is still limited, and the generalizability across diverse machines is lacking. Additionally, they also lack full consideration on the perceptual correlation between humans and machines, so their unified perceptual demands are difficult to reconcile.

This thesis regards the vision task result output by machine as a direct reflection of machine perception. Starting from the observation that machines, like humans, have perceptual limitations, the modeling methods for machine vision perceptual redundancy in image signals, the perceptual diversity among different machines, and the perceptual correlation between humans and machines are progressively proposed. Utilizing these models, the coding parameters are adaptively predicted for different image contents, alleviating the limitations of existing methods in terms of performance, generalizability, and unity. The main contribution of this thesis are as follows:

First, a modeling method for machine vision perceptual redundancy is proposed. The existence and characteristics of machine vision perceptual redundancy in image signals are studied. By introducing Just Recognizable Distortion (JRD) to measure the maximum al-

lowable coding distortion that keeps the machine visual perception consistent or correct, the relationship between machine vision perceptual redundancy and image-level quantization parameters (QP) in the coding process is established. The modeling problem of this redundancy is then transformed into the JRD prediction problem. To facilitate JRD studies, a large-scale JRD natural image dataset is constructed. Given the skewed distribution of JRD, an ensemble learning-based JRD prediction framework is proposed, which integrates multiple expert models, enabling accurate JRD prediction both with and without reference images. By adjusting coding parameters based on predicted JRD in a content-adaptive manner, the coding performance toward machine vision is improved.

Second, a modeling method for machine vision perceptual diversity is proposed. The existence of perceptual diversity among different machines and its impact on coding are studied. By introducing Satisfied Machine Ratio (SMR) to represent the proportion of machines that maintain consistent perceptions before and after coding, the modeling problem of this diversity is transformed into the SMR prediction problem. To facilitate SMR studies, by establishing representative machine libraries to annotate SMR labels, a large-scale SMR natural image dataset is constructed. Based on the nonlinear negative correlation between deep feature differences and SMR, a siamese neural network-based SMR prediction model is proposed, and a data augmentation learning strategy is designed to leverage the SMR difference between images in different distortion levels, enabling accurate SMR prediction. By adjusting coding parameters based on predicted SMR in a content-adaptive manner, the coding generalizability across diverse machines is improved.

Third, a modeling method for perceptual correlation between humans and machines is proposed. The modeling problem of this correlation is transformed into the unified prediction problem of Satisfied User Ratio (SUR) and SMR. Since existing datasets are insufficient to support efficient model training, multiple representative image quality assessment (IQA) models are aggregated to simulate the acquisition process of SUR, generating pseudo SUR labels. This enables large-scale unified pre-training via proxy task learning. Based on the multi-layer features extracted by the pre-trained model, a unified prediction network is designed. Through Difference Feature Residual Learning (DFRL), Multi-Head Attention Aggregation and Pooling (MHAAP), and spatial-channel information fusion based on MLP-Mixer, a compact feature reflecting the perceptual characteristics of both vision systems is implicitly learned, enabling accurate SUR and SMR prediction. By adjusting coding parameters based on predicted SUR and SMR in a content-adaptive manner, the coding unity to satisfied the

perceptual demands of both humans and machines is improved.

KEYWORDS: Video Coding for Machines, Perceptual Coding, Just Noticeable Distortion, Just Recognizable Distortion, Satisfied User Ratio, Satisfied Machine Ratio