2019 年 12 月 第 12 期
Vol. 40 No. 12 2019

基于智能目标检测的 HEVC 感兴趣区域编码方法

朱 威[1,2],王东洋[1],欧全林[1],郑雅羽[1,2]

[1](浙江工业大学 信息工程学院,杭州 310023)
[2](浙江省嵌入式系统联合重点实验室,杭州 310023)
E-mail: weizhu@zjut.edu.cn

摘 要: 现有的感兴趣区域编码方法主要是利用运动信息等低级视觉特征检测感兴趣区域(ROI),易受图像噪声干扰、复杂场景下的检测效果不佳并且没有检测具体内容的能力.为了能够利用高级视觉特征指导视频编码,本文提出了一种基于智能目标检测的 HEVC 感兴趣区域编码方法.首先利用深度卷积神经网络检测用户感兴趣的目标对象,然后根据检测结果确定以编码树单元(CTU)为基本单位的 ROI 区域和非 ROI 区域,再通过分析视频图像中每个像素的方向属性,进而判别 CTU 是否为平坦纹理、结构化纹理和随机纹理,并生成纹理感知图,最后对非 ROI 区域的 CTU 按纹理感知权重值进行 DCT 频率系数分级压制,以减少非 ROI 区域的码率消耗,对 ROI 区域的 CTU 按纹理感知权重下调编码量化参数(QP),以保证 ROI 区域的图像质量,从而实现智能视频编码.实验结果表明,与 HEVC 参考方法相比,本文方法在定 QP 条件下平均降低 5.67% 左右的码率;在定码率条件下,ROI 区域的 PSNR 平均提高 0.61dB,并且主观图像质量明显提升.

关 键 词: 深度卷积神经网络; HEVC; 纹理感知图; 感兴趣区域
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2019)12-2691-07

Region of Interest Coding Method for HEVC Based on Intelligent Objects Detection

ZHU Wei[1,2], WANG Dong-yang[1], OU Quan-lin[1], ZHENG Ya-yu[1,2]

[1](College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China)
[2](United Key Laboratory of Embedded System of Zhejiang Province, Hangzhou 310023, China)

Abstract: The existing region of interest coding mainly employs low-level visual features such as motion information to detect the region of interest, which is not only susceptible to image noise, and has poor inspection performances for complex scenes, but also has no ability to detect video content. In order to guide video coding with advanced visual features, this paper proposes a region of interest coding method for HEVC based on intelligent object detection. First, a deep convolutional neural network is used to detect the object of interest for the user, and then the interest region and the non-interest region with the coding tree unit(CTU) as the basic unit are determined according to the detection results. After that, by analyzing the direction attribute of each pixel in the video image, each CTU block is discriminated as a flat texture, a structural texture, or a random texture, and the texture perception map is calculated according to the texture type. Finally, the DCT coefficients of the CTUs in the non-interest region are suppressed based on the texture-aware weight to reduce the bit rate consumption, and the coding quantization parameter(QP) is adjusted down based on the texture-aware weight for the CTU in the interest region to ensure the quanlity of the interest region. Compared with the HEVC reference coding method, the proposed method reduces the bit rate by about 5.67% on average under the condition of fixed QP, and the PSNR of the region of interest increases by 0.61dB on average under the condition of fixed rate, and the subjective image quality is also significantly improved.

Key words: deep convolutional neural network; HEVC; texture perception map; region of interest

1 引 言

随着图像采集与显示技术的快速发展,高清视频已经普及,4K/8K 超高清视频正逐渐进入我们的工作和生活,视频传输与存储的数据量越来越大.新一代的视频压缩标准 HEVC 虽然较前一代的 H.264 提高了一倍左右的压缩比[1],但由于视频图像分辨率越来越高,压缩后的视频数据量仍然较大,而且目前的网络带宽资源仍然比较有限.现有的视频编码标准在对视频图像区域进行编码处理时,没有考虑人眼视觉特征,对于那些不符合人眼视觉特性的区域,却消耗不少的码率资源和计算资源.因此,如何在有限的网络带宽和存储空间下合理分配码率资源至关重要.针对上述问题,目前有效的解决方法是对视频图像中感兴趣与非感兴趣区域采用不同的编码策略,其中感兴趣区域(Region of Interest,ROI)的提出与

收稿日期: 2018-12-27 收修改稿日期: 2019-02-26 基金项目: 浙江省自然科学基金项目(LY17F010013)资助;国家自然科学基金项目(61401398)资助. 作者简介: 朱 威(通讯作者),男,1982 年生,博士,副教授,研究方向为视频编解码和智能视觉处理;王东洋,男,1990 年生,硕士研究生,研究方向为智能视觉处理;欧全林,男,1995 年生,硕士研究生,研究方向为智能视觉处理;郑雅羽,男,1978 年生,博士,副研究员,研究方向为嵌入式多媒体系统.

应用主要利用了人类视觉系统(Human Visual System,HVS)的特征[2]. HVS 在面对一个复杂视频场景时优先将注意力集中在少数具有显著视觉特征的对象上,对场景中的不同区域给于不同的关注程度[3]. 因此视频编码过程可以在 HVS 的指导下,调整感兴趣区域和非感兴趣的码率分配,提升感兴趣区域的图像质量,保证用户的视觉体验,同时降低整体压缩码率[4,5].

如何快速准确地检测和提取用户感兴趣的目标区域是实现感兴趣区域编码的重要环节,传统的提取方法主要是把运动区域当作 ROI 区域,采用帧差、光流和运动能量检测等方法虽然可以提取目标区域,但容易受运动噪声和光照等因素的影响,适用的场景有一定限制. 在最新的研究中,文献[6]针对全局运动场景的运动目标检测提出了一种基于 ORB 特征点匹配方法,首先为图像全局运动建立旋转参数模型,然后采用随机采样一致性方法筛选出最佳匹配点对,最后用帧差法得出运动目标. 文献[7]从视频码流中提取出运动矢量,对运动矢量场进行空间滤波、Mean-Shift 聚类等处理得到运动目标. 文献[8]针对高清监控视频,提出平均网格化背景建模法,该方法首先对每帧视频图像进行网格化切分,然后对网格视频帧图像运用多线程并行处理进行背景建模,最后通过鲁棒主成分分析(RPCA)方法求解提取目标对象. 文献[9]在传统视频编码框架基础上,结合人类视觉系统感知特征,根据当前编码宏块的帧间预测模式和运动矢量的大小判决 ROI 区域. 文献[10]主要针对高分辨率视频,利用视频编码得到的运动矢量信息进行权重值划分,根据相邻前景块的数量,检测出前景与背景,但整体效果不佳,运算也较为复杂. 文献[11]在压缩域中进行视频显著性检测,使用移动窗口中的离散余弦变换系数和运动信息改善视觉显著模型,并取得了一定的效果. 上述方法主要是对运动的目标选定感兴趣区域,无法知晓目标的类型,并且也不能确定静止场景下的感兴趣目标区域. 近年来,深度学习在图像分类、人脸识别等领域取得了重大突破[12],利用深度学习技术检测感兴趣的目标对象,可以提高 ROI 区域检测效果,对视频场景的适应性更好,同时支持的感兴趣目标对象更加灵活,可以较好的解决现有方法在感知灵活性和准确度上存在的问题.

根据上述分析,本文提出了一种基于智能目标检测的 HEVC 感兴趣区域编码方法. 该方法首先利用卷积神经网络提取感兴趣目标对象所在的区域; 接着对当前图像的平坦纹理区域、结构化纹理区域和复杂纹理区域进行提取,计算得到 CTU 的纹理感知权重; 然后在 HEVC 整数变换域设计自适应频率系数压制方法; 最后对非 ROI 区域频率系数进行自适应压制,对 ROI 区域调低编码量化参数,实现了比特资源合理分配,保证 ROI 区域的图像质量.

2 HVS 的感知机制分析

HVS 对视频场景中的不同区域会给予不同的关注程度,同时对视觉信号的失真也具有不同的敏感度[13]. 目前提取感兴趣目标所涉及的视觉特征和感知机制主要包括: 运动信息、人脸和肤色信息、视觉敏感度、视觉掩盖效应和中央凹感知机制[14]. 视觉感知领域根据研究内容不同可以分为低级视觉、中级视觉和高级视觉[15]. 传统的运动感知模型如图 1 所示,HVS 是按照从低级到高级的顺序对运动视觉信息进行感知[16],在大脑的初级视皮层(VI)等视觉处理区域,主要对局部运动视觉特征比较敏感,例如运动速度和运动方向; 随着视觉信号向更高级大脑区域传输,在大脑的 hMT + 等区域,HVS 会对全局性的运动视觉特征比较敏感; 在大脑的颞上沟(STS)等区域,HVS 会对特定目标的运动特征比较敏感,比如人脸的运动.

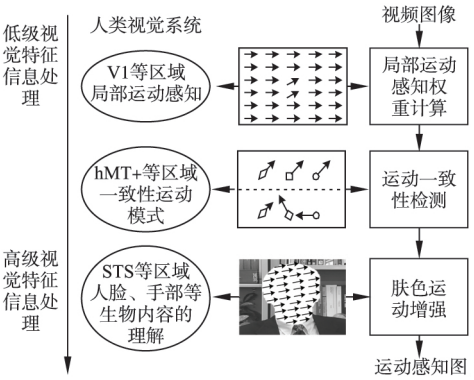

图 1 传统的运动感知模型

Fig. 1 Traditional motion perception model

虽然对于非全局运动和背景纹理简单的视频数据,利用运动视觉特征可以提取 ROI 区域,但对于全局运动场景,移动的前景对象包含的视觉信息最为丰富,HVS 对运动的前景对象具有较高的敏感度,为了提取图像区域的运动视觉特征,需要对运动对象进行检测和分割. 但现有的方法在确定运动前景对象上还存在很多不足,特别是在光照变化和摄像头运动的情况下,检测效果不佳. 除此之外,对于非运动区域,传统方法主要是利用纹理特征提取 ROI 区域,存在区域范围过大和目标不明确等问题.

高级视觉领域主要涉及的是对视频内容的识别与理解,在不同的视频场景中,人眼会有选择性的关注场景中感兴趣的目标和内容,这种现象称为 HVS 视觉注意工作机制[17]. 现有视觉注意机制主要是通过对输入视频场景进行分析,提取图像的初级视觉特征,再结合 HVS 视觉感知机制,构建出多种视觉信息的特征图,最后采用时间域或空间域特征融合的方式计算出显著性图来表示每个位置的视频显著性[18]. 利用 HVS 视觉注意模型虽然能够获取 ROI 区域,但 HVS 视觉注意模型比较复杂,只通过初级视觉特征并不能完全模拟 HVS 处理视觉信息的整个过程. 随着人工智能技术的快速发展,具有代表性的深度学习技术可以利用大量的训练样本深入地学习图像的抽象信息,更加灵活和准确地获取图像特征,实现对视频图像内容的理解和识别,为在高级视觉领域实现感兴趣区域编码提供了一种可行的途径.

3 基于智能目标检测的感兴趣区域编码方法

为了解决传统 ROI 区域编码方法在目标对象识别上的不足,提高感兴趣目标检测的灵活性,本文提出了一种基于智能目标检测的 HEVC 感兴趣区域编码方法,该方法主要应用于视频监控领域,其感兴趣目标对象的类型是由用户根据视频场景预先确定,总体流程如图 2 所示. 首先输入一帧视频图

像,利用卷积神经网络检测感兴趣目标位置,生成感兴趣目标区域; 接着根据像素的纹理方向分析当前帧的纹理复杂度,根据当前编码树单元(CTU) 的纹理复杂度计算其纹理感知权重值; 然后在已有码率控制算法和 HEVC 压缩域下,非 ROI 区域根据纹理感知权重值对 DCT 系数进行压制,减少该区域的码率资源分配,ROI 区域根据纹理感知权重值对 QP 参数值进行下调,增加 ROI 区域的码率,从而提高 ROI 区域的图像质量,实现智能视频编码.

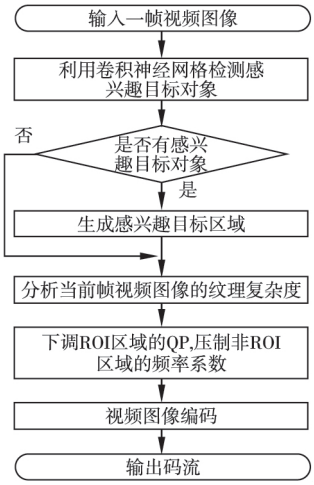

图 2　本方法流程图

Fig. 2　Flow chart of the proposed method

3.1　智能目标检测

最近几年,随着深度学习技术的快速发展,大量基于深度卷积神经网络的目标检测算法被提了出来,使得目标检测的效果取得了较大的突破. 目前常用的基于区域的目标检测方法,如 Fast-RCNN、Faster-RCNN、R-FCN 等虽然在检测精度上取得了大幅提升[19],但检测速度较慢且仅在大目标的检测效果较好. 而基于回归的目标识别方法,如 SSD、YOLO 等方法[20],采用的是端到端的目标检测与识别,在检测精度和检测速度上均获得了很大的提升,可以满足实时性的要求.

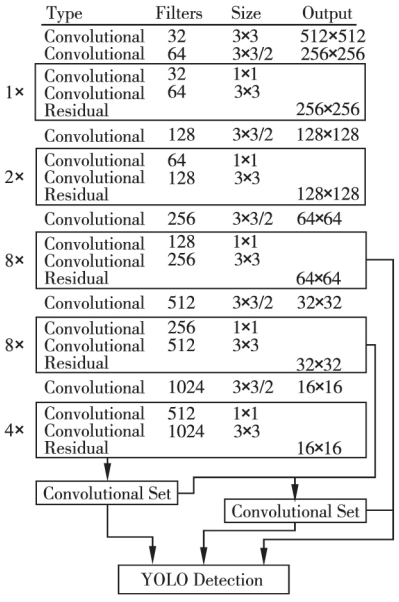

图 3　YOLO V3 网络结构

Fig. 3　Network structrue of YOLO V3

为了保证目标检测的速度和识别的准确率,并考虑到 YOLO 网络在检测速度上要优于 SSD 网络[21],因此本文方法采用 YOLO 进行感兴趣目标检测. YOLO 是一种基于回归的目标识别方法,目前已经发展到了第三代网络 YOLO V3,该网络不仅保持了 YOLO V2 的检测速度,还在小目标的检测和识别的准确率上得到了大幅提升[20],非常适合监控视频中的目标识别. YOLO V3 网络结构如图 3 所示,该网络模型使用多个表现良好的 3 × 3 和 1 × 1 卷积层,借鉴残差神经网络的思想和 Faster R-CNN 中使用的 anchor boxes 思想[22],引入多个残差网络模块,利用多尺度预测的方式改善了 YOLO V2 对小目标识别的不足.

在不同的监控视频场景中,目标对象的重要程度有所不同,因此本文方法在检测感兴趣目标对象前,由用户根据监控需求预先确定感兴趣目标对象的类型. 在视频编码过程中,将视频数据输入到 YOLO 神经网络模型中检测感兴趣目标,若当前帧检测到感兴趣目标对象,则提取所有感兴趣目标对象的位置坐标、置信度最大的目标对象类别及置信度值,为后续感兴趣区域编码提供参考; 若检测不到感兴趣目标对象,则认为不存在 ROI 区域,即整帧图像为非 ROI 区域. 将卷积神经网络应用于感兴趣目标提取,不仅可以检测运动的感兴趣目标对象,还可以检测静止的感兴趣目标对象,相比于传统的感兴趣目标检测方法,采用卷积神经网络可以提高感兴趣目标检测的灵活性. 图 4 为 Kimono 序列第 62 帧视频图像经过 YOLO 神经网络对人形目标对象经过 VOC 数据集训练之后的检测结果,矩形方框为检测框. Kimono 序列是摄像头全局移动拍摄的场景,从图 4 中可以看出,即使在全局运动的场景中卷积神经网络仍然可以准确检测到人的位置.

图 4　Kimono 序列第 62 帧目标检测结果

Fig. 4　Target detection result of the 62th frame for Kimono sequence

由于 HEVC 编码器是根据当前帧的视频图像内容自适应划分编码单元大小,一帧图像可划分多个 64 × 64、32 × 32、16 × 16 和 8 × 8 大小的编码单元(CU) ,而卷积神经网络检测出的目标区域是像素级的,因此需要对检测出的 ROI 区域边缘进行扩展处理. 根据最大 CU 块的大小,对目标区域边缘点坐标进行调整,即将检测到的目标区域矩形框上下左右四条边向外扩展到最近的 64 倍数像素边界作为 ROI 区域,其它区域为非 ROI 区域.

3.2　纹理复杂度分析

在上节获得 ROI 区域和非 ROI 区域之后,本文方法进一步对 ROI 区域和非 ROI 区域中的纹理复杂度进行分析. HVS 在关注视频场景时,一方面会对边缘方向单一的结构化纹理区域进行优先关注,而对边缘方向种类较多的复杂随机纹理区域如花草、树木等关注度较低; 另一方面,由于 HVS 视觉掩盖效应,随机纹理区域的视觉信号失真难以被发现,而结构化的纹理区域视觉信号失真具有较低的掩盖能力[23]. 因此,如何根据 HVS 对图像纹理的视觉敏感度及掩盖效应,实现纹理区域的类型划分,对实现感兴趣区域视频编码具有重要的意义. 本节通过分析图像像素的方向特性,将当前图像划分为平坦纹理区域、结构化纹理区域和复杂纹理区域,并生成纹理感知图,为后续视频图像编码提供参考. 纹理感知图的生成过程

主要包括以下三个步骤:

像素级的边缘检测. 使用四组方向不同的 5×5 高通滤波器分别计算每个像素点在 0°、45°、90°和 135°方向的边缘强度,高通滤波器模板如图 5 所示. 若每个像素点在四个方向的边缘强度都小于阈值 t_s,则认为该点不包含边缘点,否则把边缘强度最大值所对应的方向作为该点的方向属性.

CU 级的纹理复杂度检测. 统计 32×32 CU 块区域内的方向种类数及其边缘点数,若某一方向上的边缘点数大于给定的阈值 e,则认为该区域存在这个方向的纹理信息,总的方向数用 d 表示. 若总的边缘点数大于给定阈值 s,则说明边缘复杂度较高,置边缘复杂度参数 c 为 1,否则将 c 设为 0.

0°					45°					90°					135°				
0	0	0	0	0	0	0	1	0	0	0	1	0	-1	0	0	0	1	0	0
1	3	8	3	1	0	8	3	0	0	0	3	0	-3	0	0	0	3	8	0
0	0	0	0	0	1	3	0	-3	-1	0	8	0	-8	0	-1	-3	0	3	1
-1	-3	-8	-3	-1	0	0	-3	-8	-1	0	3	0	-3	0	0	-8	-3	0	0
0	0	0	0	0	0	0	-1	0	0	0	1	0	-1	0	0	0	-1	0	0

图 5 四个方向上的高通滤波器

Fig. 5 High-pass filters under four directions

CTU 级的感知图生成. 将权重值设置为高、中、低三档,对应的数值分别为 2、1 和 0,首先根据 32×32 CU 块中的纹理方向总数 d 计算该像素块的纹理感知权重值 T_{32},如式(1)所示. 若 d 值为 0,表示当前像素块中无明显方向,纹理比较平坦,HVS 对该区域具有中等敏感度,感知权重值设为 1; 若 d 值为 1,表示当前像素块中只有一个方向,是比较明显的结构化纹理,HVS 对该区域具有较高的敏感度,感知权重值设置 2; 若 d 值为 2,表示当前块中有两个显著纹理方向,在高感知权重的基础上使用边缘复杂度参数 c 进行调整,c 为 1 则感知权重降为 1; 若 d 值为 3,表示当前块中有三个明显方向,方向数较多,因此在中感知权重的基础上同样使用边缘复杂度参数 c 进行调整; 若 d 值大于 3,表示当前块中包含的方向数较多,内部很有可能为随机性纹理,HVS 对该区域敏感度比较低,因此将感知权重设为最小值 0.

$$T_{32}=\begin{cases}1, & \text{if } d=0\\ 2, & \text{if } d=1\\ 2-c, & \text{if } d=2\\ 1-c, & \text{if } d=3\\ 0, & \text{others}\end{cases} \tag{1}$$

T_{32}的大小是以 32×32 像素块为单位的,为了得到 CTU 级的 64×64 像素块大小的纹理感知图 T_{64},需要对 T_{32}进行后处理操作. 首先统计每个 64×64 像素块中的四个 32×32 像素块的纹理复杂度,参数 z 和 t 分别表示感知权重值为 0 和 2 的 32×32 像素块个数,然后按式(2)得到 T_{64}.

$$T_{64}=\begin{cases}t!=0?\ 2:1, & \text{if } z=0\\ t==3?\ 2:1, & \text{if } z=1\\ t==2?\ 1:0, & \text{if } z\geqslant 2\end{cases} \tag{2}$$

图 6 为 Kimono 序列第 62 帧原始视频图像经过纹理复杂度分析生成的纹理感知图. 从图 6 中可以看出,以检测框为分界线,非 ROI 区域和 ROI 区域中不同的灰度值代表不同的纹理感知权重,灰度值越亮的区域敏感度越高. 图 6 中大部分纹理复杂的区域(背景的松树叶)都识别为较低的感知权重区域,而边缘方向单一的区域(前景的人物)都被识别为较高的感知权重区域,检测结果能够较好地反映视频图像中各个区域的纹理感知效果.

图 6 Kimono 序列第 62 帧纹理感知图

Fig. 6 Texture perception map of the 62th frame for Kimono sequence

3.3 频率系数压制矩阵

HEVC 继承了传统视频编码中的预测残差 DCT 变换方法,经 DCT 变换后,绝大部分能量都集中于矩阵左上角的低频系数中,图像中较多的细节信息会分散在高频区域. 考虑到 HVS 对高频信号的失真敏感度较低,本文方法在 HEVC 变换域上,对人眼视觉敏感度较低的区域进行较高强度的频率系数压制,对人眼视觉敏感度较高的区域采用较低强度的频率系数压制或不进行压制,整体压制策略如式(3)所示:

$$\bar{Y}=Y\otimes S_n \tag{3}$$

式中,$\bar{Y}$ 表示压制后的频率系数矩阵,$\otimes$表示两个矩阵对应元素相乘,S_n 表示频率系数压制矩阵,一般形式如式(4)所示,n 为矩阵大小,与 HEVC 变换单元(TU)的大小一致,取值为 4、8、16 和 32,矩阵 S_n 中的元素取值为 0 或 1,并且满足 $s_{n+1}\leqslant s_n$ 的约束条件,从而保证了频率系数压制是从高频到低频过渡的过程.

$$S_n=\begin{bmatrix} s_1 & s_2 & s_3 & \cdots & s_n\\ s_2 & s_3 & \cdots & s_n & \cdots\\ s_3 & \cdots & s_n & \cdots & s_{2n-3}\\ \cdots & s_n & \cdots & s_{2n-3} & s_{2n-2}\\ s_n & \cdots & s_{2n-3} & s_{2n-2} & s_{2n-1}\end{bmatrix} \tag{4}$$

在频率系数压制时,本文方法使用三种频率系数压制矩阵实现从高频到低频的过渡,可根据编码块的视觉重要程度选择不同级别的 S_n 进行压制. 针对 4×4、8×8、16×16 和 32×32 变换块设置了三种候选频率系数压制矩阵,候选频率系数压制矩阵组按式(5)计算得到.

$$S_n(k)_{ij}=\begin{cases}1, \text{if } (i+j)\leqslant \dfrac{n}{4}\times k+m\\ 0, \text{others}\end{cases} \tag{5}$$

式中 i 和 j 分别为矩阵元素的横纵坐标,取值范围都为 $[0,n-1]$,k 为 $n\times n$ 大小的 3 种压制矩阵的索引,取值为 1、2 和 4,压制强度依次增强,m 为偏移量,本文取值为 0; 其中以 8×8 块压制矩阵为例,3 种候选频率系数的压制矩阵组如图 7 所示.

$S_8(1)$								$S_8(2)$								$S_8(4)$							
1	1	0	0	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1
1	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1	1	1	1	1	0
0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0

图 7 8×8 候选压制矩阵组

Fig. 7 8×8 Candidate suppression matrixs

3.4 基于 ROI 的编码策略

现有编码方法中的码率控制技术主要是为了控制码率大小而进行比特资源分配,没有考虑到不同区域的视觉差异. 为了优

先保证 ROI 区域的图像质量,本文根据 ROI 区域的纹理感知权重,对 ROI 区域 QP 值进行不同程度的下调,对于每个待编码 CTU,根据其纹理感知权重值 T,计算其 QP 参数下调值 DQP,如式(6)所示. 若 T_{64} 等于 0,则表示当前 CU 为随机纹理区域,DQP 取值为 2; 若 T_{64} 等于 1,则表示当前 CU 为平坦区域,DQP 取值为 4; 若 T_{64} 等于 2,则表示当前 CU 为结构化纹理区域,DQP 取值为 6. 由于 I 帧只采用帧内预测,其平均编码字节数是 P 帧的数倍,为了不进一步加重峰值码率,本方法不对其 QP 进行下调.

$$DQP=\begin{cases}2, \text{if } T_{64}=0\\4, \text{if } T_{64}=1\\6, \text{if } T_{64}=2\end{cases} \tag{6}$$

由于 HVS 对非 ROI 区域的关注程度不高,本文方法对非 ROI 区域进行频域系数压制. 对于非 ROI 区域每个 CTU 根据其纹理感知权重值选择频率系数压制矩阵对其 DCT 频率系数进行不同程度的压制,对随机纹理区域进行高强度压制,对平坦区域进行中等强度压制,对结构化纹理区域进行较低强度压制,具体压制方法如式(7)所示. 若 T 为 0,则选择 $S_n(1)$ 对随机纹理区域进行较强压制; 若 T 为 1,则选择 $S_n(2)$ 对平坦区域进行中等强度压制; 若 T 为 2,则选择 $S_n(4)$ 对结构化纹理区域进行较弱强度压制. 由于 I 帧是后续 P 帧的参考基础,如果 I 帧失真后面的 P 帧将会受到影响,为了保证整体视频图像的质量,本方法不对 I 帧的非 ROI 区域进行系数压制. 此外,为了进一步减少预测误差扩散,本文方法只对奇数帧进行压制,即进行隔帧压制,减少非 ROI 区域压制对 ROI 区域图像质量影响.

$$W_n=\begin{cases}S_n(1), \text{if } T=0\\S_n(2), \text{if } T=1\\S_n(4), \text{if } T=2\end{cases} \tag{7}$$

4　实验与分析

4.1　实验环境及配置

本实验采用 YOLO V3 神经网络进行智能目标检测,模型训练和测试所使用的软硬件平台如下: OS: Ubuntu 16.04 LTS; CPU: Intel Core i7-8700K CPU @ 3.70GHz; GPU: NVIDIA GeForce GTX1080 Ti x 2; 内存: 32G. 由于人是最为常见的监控目标,本文选择的感兴趣目标对象类型为人形目标,并选择 VOC2012 数据集中的人形目标进行训练.

本实验 HEVC 编码软件为 X265_1.8,开发环境为 Visual Studio2012,测试平台的处理器为 Intel Core i5-2520,主频 2.5GHz. X265 编码器的配置如下: 帧率 30fps,IPP 模式,I 帧间隔为 100,DCT 系数压制采用奇数帧压制. 实验选取了四个全高清的 HEVC 参考视频序列 Kimono、BasketballDrive、Poznan_CarPark 和 Tennis 验证本文方法的有效性,每个序列选取前 100 帧进行统计分析. 实验中,本文方法中的阈值 t_s 设为 3,阈值 e 设为 100,阈值 s 设为 400,使用 VOC2012 数据集训练好的 YOLO V3 网络模型对上述序列进行检测并输出检测结果,用于提取后续的 ROI 区域.

4.2　检测模型训练及结果

在模型的训练过程中,初始学习率设为 0.01,衰减系数设置为 0.00050,训练集的目标置信度设为 0.5. 为了防止过拟合现象,训练阶段采用动量为 0.9 的异步随机梯度下降,实验训练次数为 100000 次. 为了测试最佳的权重文件,训练时采用每 1000 次迭代保存一次权重文件. 为了提高检测精度,将训练图像的分辨率从默认的 416×416 提高到 618×618,训练集使用 VOC2012-trainval 数据集,测试集使用 VOC2017-test 数据集. 将 YOLO V3 网络在数据集上进行测试,测试集的目标置信度设为 0.25,最终计算得出准确率为 0.84,召回率为 0.80. 图 8 为训练冻结之后 YOLO 对测试集的检测结果,从图中可以看出 YOLO V3 网络能够对不同尺度的人形目标进行有效地检测.

图 8　YOLO V3 对测试集的检测结果

Fig. 8　Detection result of YOLO V3 for the test set

4.3　实验结果与分析

本实验使用两种方法对所选的视频序列进行编码: 第一种为 X265 参考编码方法; 第二种为本文提出的感兴趣区域编码方法. 这两种方法分别在固定 QP 和固定码率条件下进行实验对比,其中 QP 分别设置为 24、27、30,目标码率分别设置为 2048 kbps、4096 kbps、6144 kbps. 固定 QP 条件下的测试主要是衡量本文方法中频域系数压制部分在降低码率方面的性能,固定码率条件下的测试主要是衡量本文方法整体的处理效果. 通过比较本文方法相对于参考方法对感兴趣区域的 PSNR 增益,可以衡量本文方法的编码效果.

表 1 为本文方法与参考方法在固定 QP 条件下的性能对比. 从表中可以看出,相比于参考方法,本文方法的整帧 PSNR 平均降低 0.32dB,ROI 区域的 PSNR 平均仅降低了 0.11dB,而实际输出码率平均减少了 5.67%. 因此,从降低码率的角度来说,本文方法中的频率系数压制部分在保证 ROI 区域图像质量降低较小的情况下,节省较多的比特资源. 图 9 为 Tennis 序列在 QP 为 24 的配置下第 30 帧两种方法的重建

(a)参考方法　　(b)本文方法

图 9　Tennis 序列第 30 帧下的定 QP 主观质量比较(QP=24)

Fig. 9　Comparison of subjective quality under QP24 for the 30th frame of Tennis sequence

图像主观质量对比. 对于该序列,本文方法相对于参考方法的 ROI 区域 PSNR 平均仅降低了 0.014dB,整帧 PSNR 平均降低了 0.696dB,消耗的平均码率减少了 7.17%. 从图 9 中可以看出,经本文方法编码后图像的非 ROI 区域(背景区域)与参考方法编码后图像的主观质量差异较小,对于 ROI 区域,也就

是运动员所在的区域,本文方法和参考方法视觉感知质量基本一致. 因此在固定 QP 条件下,虽然本文方法编码的整帧图像的 PSNR 相对于参考方法略有下降,但由于引起 PSNR 下降部分的区域主要为非 ROI 区域,从降低码率角度而言本文方法在保证 ROI 区域信息失真较小的情况下,可以节省较多的比特资源. 由于本文方法采用的是隔帧进行压制,即使压制帧出现明显的视频失真,后一帧在编码时也会将出现的失真进行改善,并且两帧之间的时间间隔较短,能够对失真进行掩盖. 此外,本文方法主要是针对高频分量进行压制,对图像的主观质量影响较小,这符合人眼对视觉感知编码的实际需求.

表 1 本文方法与参考方法在固定 QP 下的性能比较

Table 1 Performance comparison between the proposed method and the anchor method under fixed QPs

序列	QP	本文方法			参考方法			ΔPSNR_ROI (dB)	ΔPSNR (dB)	ΔBit (%)
		实际码率 (kbps)	整帧图像 PSNR (dB)	ROI 区域 PSNR (dB)	实际码率 (kbps)	整帧图像 PSNR (dB)	ROI 区域 PSNR (dB)			
Kimono	24	8540.13	41.479	41.78	8954.68	41.761	41.776	0.004	−0.282	−4.63%
	27	5673.91	40.645	40.77	5858.36	40.853	40.77	0	−0.208	−3.15%
	30	4015.76	39.065	39.104	4059.12	39.734	39.587	−0.483	−0.669	−1.07%
Basketball Drive	24	12591.19	39.016	38.711	13567.05	39.316	38.712	−0.01	−0.3	−7.19%
	27	6775.07	38.053	37.337	7090.36	38.251	37.338	−0.01	−0.198	−4.45%
	30	4189.32	37.124	36.246	4321.63	37.266	36.243	0.003	−0.142	−3.06
Poznan _CarPark	24	3232.71	39.969	39.728	3704.64	40.191	40.09	−0.362	−0.222	−12.74%
	27	1802.3	39.054	38.236	1974.56	39.205	38.471	−0.235	−0.151	−8.72%
	30	1196.29	38.101	36.758	1279.95	38.209	36.913	−0.155	−0.108	−6.54%
Tennis	24	10902.54	39.946	40.87	11744.49	40.642	40.884	−0.014	−0.696	−7.17%
	27	7118.28	38.993	39.695	7504.99	39.513	39.711	−0.016	−0.52	−5.15%
	30	4846.89	37.933	38.434	5058.4	38.32	38.457	−0.023	−0.387	−4.18%
Average		5907.03	39.11	38.97	6259.85	39.44	39.08	−0.11	−0.32	−5.67%

表 2 本文方法与参考方法在固定码率下的性能比较

Table 2 Performance comparison between the proposed method and the anchor method under fixed bit rates

序列	目标码率 (kbps)	本文方法			参考方法			ΔPSNR_ROI (dB)	ΔPSNR (dB)	ΔBit (%)
		实际码率 (kbps)	整帧图像 PSNR (dB)	ROI 区域 PSNR (dB)	实际码率 (kbps)	整帧图像 PSNR (dB)	ROI 区域 PSNR (dB)			
Kimono	2048	1981.24	36.33	37.113	1895.7	36.632	36.44	0.673	−0.302	4.51%
	4096	3974.34	38.795	39.482	3754.68	39.266	39.013	0.469	−0.471	5.85%
	6144	6001.14	39.95	40.671	5632.19	40.533	40.347	0.32	−0.583	6.55%
Basketball Drive	2048	2173.32	34.992	34.266	2097.19	35.119	33.772	0.494	−0.127	3.63%
	4096	4363.09	36.884	36.255	4263.7	37.145	35.898	0.357	−0.261	2.33%
	6144	6460.28	37.579	37.049	6413.25	37.974	36.807	0.242	−0.395	0.73%
Poznan _CarPark	2048	1751.89	37.412	37.546	1719.11	37.48	36.985	0.561	−0.068	1.91%
	4096	3790.61	39.318	39.585	3749.4	39.408	39.058	0.527	−0.09	1.10%
	6144	5811.28	40.01	40.266	5778.12	40.086	39.699	0.567	−0.076	0.57%
Tennis	2048	2049.61	34.665	35.381	1959.12	34.748	34.33	1.051	−0.083	4.62%
	4096	4129.74	37.108	38.12	3947.17	37.275	37.055	1.065	−0.167	4.63%
	6144	6182.17	38.271	39.459	5953.78	38.623	38.509	0.95	−0.352	3.84%
Average		4055.7	37.61	37.93	3930.28	37.86	37.33	0.61	−0.25	3.36%

表 2 为本文方法和参考方法在固定码率配置下的性能对比. 从表 2 中可以看出四个视频序列在设置的固定码率下本文方法相对于参考方法的 ROI 区域平均 PSNR 增益达到 0.61dB,整帧平均 PSNR 仅减少了 0.25dB. 其中 Tennis 序列的 ROI 区域编码图像质量改善最为显著,这主要是因为该视频序列的非 ROI 区域包含大量的复杂纹理区域,而参考方法在编码这部分区域消耗了较多的比特资源,而本文方法根据 ROI 区域进行比特资源优化,获得了较好的编码效果. Kimono、BasketballDrive 和 Poznan_CarPark 这三个序列的 ROI 区域 PSNR 的提升也较为明显,这是因为这三个序列的非 ROI 区域的面积较大,并且背景视频信号的噪声也较强,可以节省较多的比特资源分配给 ROI 区域,进而提升 ROI 区域的图像质量. 图 10 为本文方法与参考方法在 Kimono 序列第 11 帧的编码重建图像主观质量对比,从图 10 中可以看出,采用本文方法编码后获得的 ROI 区域主观图像质量要明显好于参考方法: 女士的衣服、头发、眼角、下巴等部位更为清晰. 从上述实验数据可以看出,相对于参考方法,本文方法的 ROI 区域 PSNR 得到了提升,整帧 PSNR 相对于参考方法有所下降,但

由于引起 PSNR 下降的区域是属于视觉不重要的区域,对主观视觉的影响较小. 因此从整体效果来看采用本文方法有效地改善了视频图像的视觉感知效果.

(a)参考方法　(b)本文方法

图 10　Kimono 序列第 11 帧下的定码率主观图像质量比较(2048 kbps)

Fig. 10　Comparison of subjective quality under 2048 kbps for the 11th frame of Kimono sequence

5　结　论

本文利用深度学习视觉检测技术,提出了一种基于智能目标检测的 HEVC 感兴趣区域编码方法. 首先通过卷积神经网络检测视频图像中感兴趣目标,生成 ROI 区域; 接着通过分析像素级的方向属性生成纹理感知图; 最后利用纹理感知图,对非 ROI 区域的 DCT 频率系数进行多级压制,对 ROI 区域的 QP 值进行不同程度地下调,在已有码率控制框架的基础上,保证了 ROI 区域的图像质量,减少非 ROI 区域的码率资源消耗,从而实现智能视频编码. 与传统的 ROI 编码方法相比,本文方法对感兴趣目标检测方面具有更好的灵活性,弥补了传统方法在 ROI 区域提取过程使用初级视觉特征的局限性,编码后的重建图像更加符合 HVS 的高级视觉感知要求. 当然本文方法还存在一些问题,尤其是场景中存在较多的感兴趣目标对象时,处理效果还需要提升,我们将在后续的工作中进一步研究和完善.

References:

[1] Bossen F,Bross B,Suhring K,et al. HEVC complexity and implementation analysis [J]. IEEE Transactions on Circuits & Systems for Video Technology,2013,22(12):1685-1696.

[2] Bari A,Robbins T W. Inhibition and impulsivity: behavioral and neural basis of response control [J]. Progress in Neurobiology,2013,108(9):44-79.

[3] Zebbiche K,Khelifi F. Efficient wavelet-based perceptual watermark masking for robust fingerprint image watermarking [J]. IET Image Processing,2014,8(1):23-32.

[4] Wang S,Ma S,Wang S,et al. Rate-GOP based rate control for high efficiency video coding [J]. IEEE Journal of Selected Topics in Signal Processing,2013,7(6):1101-1111.

[5] Wang Kai,Wu Min,Yao Hui,et al. Moving target detection method based on multi-frame background subtraction and double threshold [J]. Journal of Chinese Computer Systems,2017,38(1):179-183.

[6] Li Xiao-hong,Xie Cheng-ming,Jia Yi-zhen,et al. Rapid moving object detection algorithm based on ORB features [J]. Journal of Electronic Measurement and Instrument,2013,27(5):455-460.

[7] Sun Le,Dai Ming,Li Gang,et al. An algorithm of mean-shift clustering-based moving object segmentation in H. 264 compression domain [J]. Journal of Optoelectronics Laser,2013,24(11):2205-2211.

[8] Wang Zhou,He Jun,Hu Zhao-hua. Multi-ROI background modeling method suitable for HD monitor video [J]. Journal of Chinese Computer Systems,2018,39(6):1190-1194.

[9] Liu Peng-yu,Jia Ke-bin. Fast extraction and coding algorithm for video region of interest [J]. Journal of Circuits & Systems,2013,18(2):413-419.

[10] Praeter J D,Vyver J V D,Kets N V,et al. Moving object detection in the HEVC compressed domain for ultra-high-resolution interactive video [C]//IEEE International Conference on Consumer Electronics,2017:135-136.

[11] Wang H,Wang L,Hu X,et al. Perceptual video coding based on saliency and just noticeable distortion for H. 265/HEVC [C]//International Symposium on Wireless Personal Multimedia Communications,2014:106-111.

[12] Lecun Y,Bengio Y,Hinton G. Deep learning [J]. Nature,2015,521(7553):436-521.

[13] Yu H,Chang Y,Lu P,et al. Efficient object detection based on selective attention [J]. Computers & Electrical Engineering,2014,40(3):907-919.

[14] Yoshimoto S,Uchida-Ota M,Takeuchi T. Effect of light level on the reference frames of visual motion processing [J]. Journal of Vision,2014,14(13):1-28.

[15] Rosen M L,Stern C E,Michalka S W,et al. Cognitive control network contributions to memory-guided visual attention [J]. Cerebral Cortex,2015,26(5):2059-2073.

[16] Fernandino L,Binder J R,Desai R H,et al. Concept representation reflects multimodal abstraction: a framework for embodied semantics [J]. Cerebral Cortex,2015,26(5):2018-2034.

[17] Mackenzie A K,Harris J M. Eye movements and hazard perception in active and passive driving [J]. Visual Cognition,2015,23(6):736-757.

[18] Yang R,Xu M,Wang Z,et al. Saliency-guided complexity control for HEVC decoding [J]. IEEE Transactions on Broadcasting,2018,64(4):865-882.

[19] Chen L C,Papandreou G,Kokkinos I,et al. DeepLab: semantic image segmentation with deep convolutional nets,atrous convolution,and fully connected CRFs [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2018,40(4):834-848.

[20] Redmon J,Divvala S,Girshick R,et al. You only look once: unified,real-time object detection [C]//IEEE Computer Vision and Pattern Recognition,2016:779-788.

[21] Redmon J,Farhadi A. YOLO9000: better,faster,stronger [C]//IEEE Conference on Computer Vision & Pattern Recognition,2017:6517-6525.

[22] Ren S,He K,Girshick R,et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2015,39(6):1137-1149.

[23] Jiang Gang-yi,Huang Da-jiang,Wang Xu,et al. Overview on image quality assessment methods [J]. Journal of Electronics & Information Technology,2010,32(1):219-226.

附中文参考文献:

[5] 王　凯,吴　敏,姚　辉,等. 多帧背景差与双门限结合的运动目标检测方法 [J]. 小型微型计算机系统,2017,38(1):179-183.

[6] 李小红,谢成明,贾易臻,等. 基于 ORB 特征的快速目标检测算法 [J]. 电子测量与仪器学报,2013,27(5):455-460.

[7] 孙　乐,戴　明,李　刚,等. H. 264 压缩域中 mean-shift 聚类运动目标分割算法 [J]. 光电子·激光,2013,24(11):2205-2211.

[8] 汪　舟,何　军,胡昭华. 适宜于高清监控视频的多 ROI 背景建模方法 [J]. 小型微型计算机系统,2018,39(6):1190-1194.

[9] 刘鹏宇,贾克斌. 视频感兴趣区域快速提取与编码算法 [J]. 电路与系统学报,2013,18(2):413-419.

[23] 蒋刚毅,黄大江,王　旭,等. 图像质量评价方法研究进展 [J]. 电子与信息学报,2010,32(1):219-226.