高技术通讯 2023 年 第33 卷 第10 期:1068-1076

doi:10.3772/j.issn.1002-0470.2023.10.007

基于 T-CNN 的 3D-HEVC 深度图帧内快速编码算法①

于 源② 贾克斌③

(北京工业大学信息学部 北京 100124)

(北京工业大学计算智能与智能系统北京市重点实验室 北京 100124)

(先进信息网络北京实验室 北京 100124)

摘 要 3D-HEVC 标准中引入了具有大面积平坦区域、陡峭边缘和低纹理复杂度特性的深度图。针对深度图编码过程中编码单元(CU)率失真优化导致编码复杂度过高这一问题,本文在分析深度图编码所具有的特点的基础上,构建了深度图划分深度数据集,并提出了一种基于两通道特征传递卷积神经网络(T-CNN)的划分深度预测算法。使用本文提出的算法替换原始编码器中各视点下深度图 CU 划分模块,可以在一定的率失真性能损失下,将原始 HTM-16.0 编码器编码时间平均减少 76% 左右,编码效率得到了显著提升。

关键词 3D-HEVC;深度图;帧内编码;卷积神经网络

0 引 言

目前,在日常生活中广泛应用的视频编码标准是 H.264/AVC(advanced video coding)标准[1]。然而,在有限的网络带宽和存储资源下,该标准已逐渐难以满足高分辨率视频业务对于高效率编码的要求。为此,新一代编码标准 H.265/HEVC(high efficiency video coding)[2]应运而生。由于在二维视频编码中 H.265/HEVC 标准展现了其高效性,因而基于 H.264/AVC 的三维视频编码标准(three dimensional AVC,3D-AVC)也发展到了基于 H.265/HEVC 的 3D-HEVC 标准,该标准是目前最新一代的 3D 视频编码标准。

以往的多视点视频编码(muliti-view video coding,MVC)[3]不包括深度信息,因而不能使用基于深度图的绘制技术(depth image based rendering,DIBR)[4]来合成虚拟视点,相比之下加入了深度图的 3D-HEVC 标准的显示效果更好且范围更广。然而,3D-HEVC 标准需要对多个视点中的纹理图和对应深度图进行编码,这导致数据量急剧增加。该标准下各个视点中的纹理图和深度图仍以 HEVC 标准的编码框架为基础[5]。H.265/HEVC 的编码复杂度相对于 H.264/AVC 增加了 253%,其中基于四叉树结构的编码单元(coding unit,CU)[6]递归划分技术是编码复杂度提升的主要源头,判断其划分过程就占据了整体编码时间的 80%[7]左右。3D-HEVC 继承了 HEVC 中的这种划分结构,且所有视点中的纹理图和深度图均需要进行该种划分过程。在这个过程中,从最小尺寸 8×8 的 CU 到最大尺寸 64×64 的 CU 中所有可能的划分方式均要先计算率失真成本(rate-distortion cost,RDCost),随后选取 RDCost 值最低的划分方式为当前编码树单元(coding tree unit,CTU)的最终划分结构,对当前帧的所有 CU 进行处理后即可得到该帧的最终划分结果。在待编码 CU 的划分过程中,需要在多达 35 种 HEVC 原有的帧内预测模式以及 3D-HEVC 中加入的深度建模模式(depth modeling mode, DMM)DMM1 和 DMM4 中

① 北京市自然科学基金(4212001)资助项目。

② 女,1997 年生,硕士生;研究方向:三维视频编码技术;E-mail: yuyuan1119@163.com。

③ 通信作者,E-mail: kebinj@bjut.edu.cn。

(收稿日期:2022-09-22)

依据 RDCost 进行最优帧内模式的选择,大幅提高了编码复杂度。本文也将从这一点出发,提出优化算法来加快深度图帧内 CU 划分过程,提高编码效率。

现有的降低 3D-HEVC 帧内编码复杂度的研究可主要分为 2 类,包括预测模式快速决策算法和 CU 尺寸快速决策算法。文献[8]使用 Canny 算子和 Hough 变换处理深度图中所具有的独特的边缘信息,最终跳过计算复杂度高的 DMM1 模式,加快帧内预测模式选择过程。此外,各向同性的 Sobel 算子也可以被用来检测预测单元的纹理复杂度和边缘方向[9]。

由于帧内 CU 划分具有的特点,机器学习相关的方法也广泛应用到降低 3D-HEVC 深度图编码计算复杂度的工作中。比较有代表性的算法如基于静态决策树的快速深度图编码算法[10]和利用多个决策树进行帧内编码单元划分深度早期决策的算法[11]。但是,上述方法都是基于概率或人工特征判断,缺乏鲁棒性。近些年深度学习快速发展,有效克服了传统机器学习方法中存在的这些弊端。如文献[12]中使用 10 层的快速选择卷积神经网络(fast selecting CU's depth-convolutional neural network, FSCD-CNN)对深度图 CU 的分类进行学习,加快视频编码的速度。文献[13]使用整体嵌套边缘检测(holistically-nested edge detection, HED)[14]网络来检测深度图的边缘,通过对边缘复杂度进行提前判断来简化划分深度选择过程,实现快速编码。为了加快所有视点下深度图的编码过程,本文使用所提出的算法对深度视频进行 CU 划分深度预测,并使用其来替换原始编码器 HTM 中的深度图 CU 划分过程,显著地提高了编码效率。本方法的参数量相比复杂的网络更低,计算时间也较少,可以在未来更加易于部署到硬件中实现。

1　3D-HEVC 标准编码结构

1.1　3D-HEVC 编码流程

如图 1 所示,3D-HEVC 标准中同一时刻的所有视点的纹理图及深度图组成一个处理单元(access unit,AU),并以此作为单位按照时间顺序进行编码。为了保证编码质量,独立视点按照原始的 HEVC 标准进行编码,依赖视点采用扩展后的 HEVC 标准进行编码。该扩展过程中加入了更加适用于深度图编码以及多视点视频编码的新技术,如基于深度图特点提出了新的帧内预测模式,深度建模模式 DMMs 等,然而新模式的加入使得深度图帧内 CU 划分的复杂度再次增加。

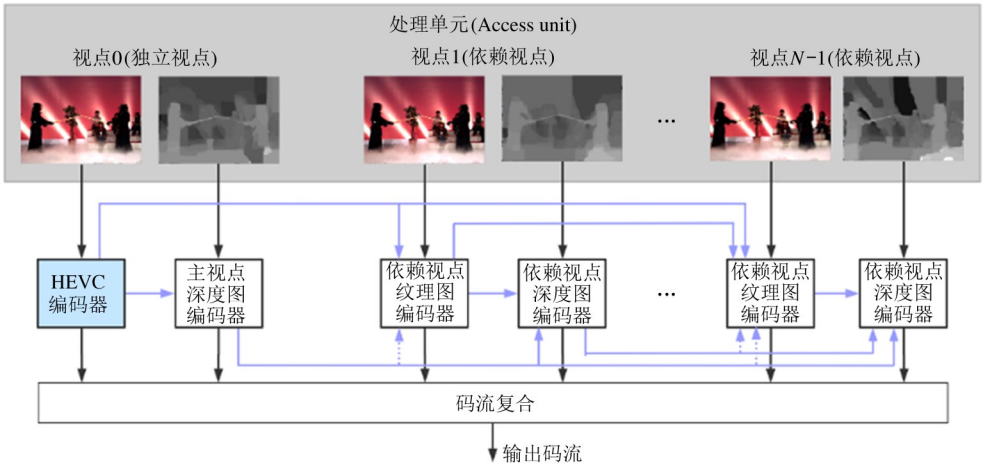


图 1　3D-HEVC 多视点标准编码顺序

1.2　CTU 划分结构

其中,占据编码复杂度最高的 CTU 划分过程的具体划分方式如图 2 所示。

3D-HEVC 标准下的帧内 CU 采用四叉树结构进行划分。如图 2 所示,待编码的图像以 CTU 为单位进行划分,默认情况下 CTU 的尺寸为 64×64,该尺寸的 CTU 称作最大编码单元(largest coding unit, LCU)。CTU 可以包含单个 CU,即 CTU 不再进一步

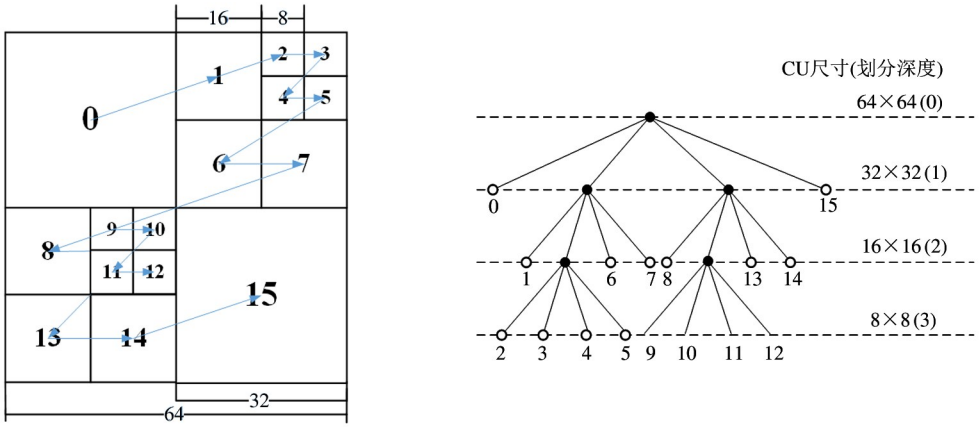


图 2　CTU 划分结构示意图

划分,也可以根据四叉树结构递归拆分成多个较小的 CU,如图中不同划分深度下不同尺寸的 CU,最小尺寸默认值为 8×8。

每个 CTU 中的 CU 大小是通过蛮力率失真优化搜索来进行确定的,包括从父 CU 到子 CU 由上而下的检查过程以及由子 CU 到父 CU 的比较过程。在检查过程中,编码器需要检查整个 CTU 的率失真代价,随后对其子 CU 进行检查,此过程由上到下进行,直到 CU 尺寸达到最小。父 CU 的率失真代价使用R^{parent}表示,其子 CU 率失真代价表示为 $R_i^{\text{child}}(i \in \{1,2,3,4\})$。根据父 CU 和其对应的子 CU 的率失真代价,进行由下至上的比较过程来判断是否拆分父 CU。若满足 $\sum_{i=1}^{4} R_i^{\text{child}} \geqslant R^{\text{parent}}$,则父 CU 不进行拆分;若满足 $\sum_{i=1}^{4} R_i^{\text{child}} < R^{\text{parent}}$,则父 CU 将被拆分。在决定是否拆分时,要考虑划分标志的率失真代价。在经过完整的率失真优化搜索后,最终率失真代价最小的 CU 划分结构将会被采纳。

1.3　深度图编码特性

从图 3 以及图 4 中的数据统计结果可以看出,深度图中包含大面积的平坦区域,这使得 50% 左右的 CU 的划分深度为 0,即所有待划分的 CTU 中约一半是不需要进行划分的,然而这些不需划分的 CU 在标准编码器HTM中仍要进行率失真成本的计算

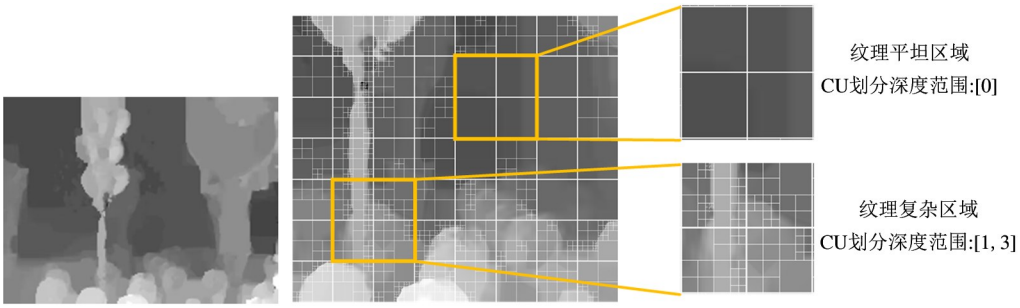


图 3　深度图特性

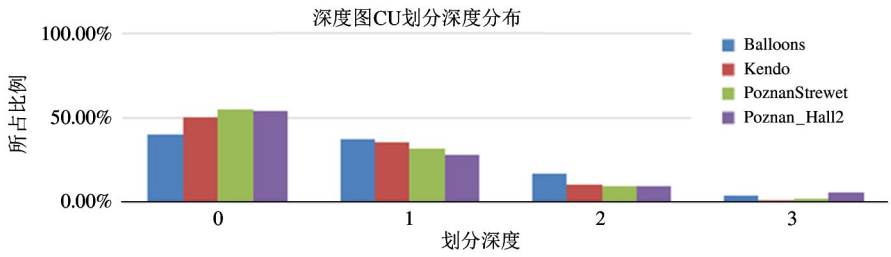


图 4　深度图划分深度统计

等不必要的复杂操作,这就导致深度图编码时间急剧增加。从图 5 中可以看到,在不同的量化参数(quantization parameter,QP)值下,深度图编码时间占总编码时间的 86% ~88%,即基于 3D-HEVC 的多视点编码过程中,深度图的编码占据了绝大多数编码时间,因而急需对深度图编码过程进行优化。

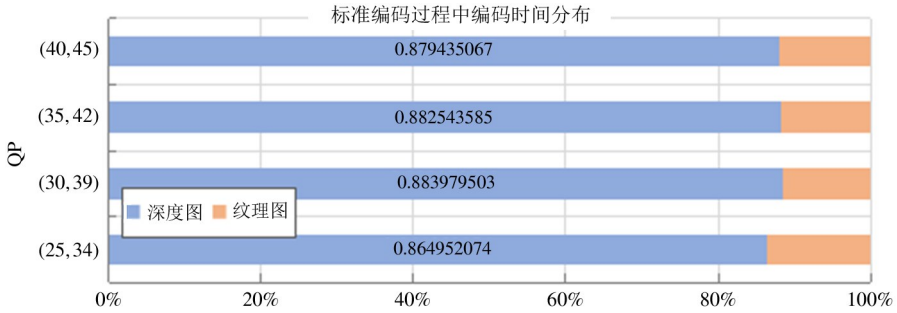


图 5　编码时间分布

2　算法设计

2.1　深度图划分数据集

表 1 中所示的是构建深度图划分数据集所使用的视频及各项参数。构建的数据集将用于后续的网络训练。由于 3D-HEVC 标准测试序列数量有限,为了尽量扩大数据集的数量和种类,选择的视频数量较多,且为了避免训练数据和测试数据出现重叠,将训练帧和测试帧以至少 50 帧完全间隔开。

表 1　深度划分数据集

视频序列	分辨率	编码帧	数据集构建帧	选择的视点
Balloons	1024 ×768	0 ~50	100 ~300	5
UndoDancer	1920 ×1088	0 ~50	100 ~250	9
Outdoor	1024 ×768	仅训练	0 ~100	10
Lovebird	1024 ×768	仅训练	0 ~240	4
Book	1024 ×768	仅训练	0 ~100	10
Shark	1920 ×1088	仅训练	100 ~300	1
Gt _ Fly	1920 ×1088	仅训练	100 ~250	1
PoznanCarpark	1920 ×1088	仅训练	100 ~250	5

2.2　T-CNN 网络结构

图 6 展示的是本文的整体算法流程。由于深度图中包含大面积平坦区域以及分割平坦区域的边缘部分,因而本文选择搭建 2 个通道的特征传递层来更有效地提取特征。在预处理部分,将深度视频的待编码帧裁剪成 64 ×64 尺寸的 LCU,传入网络。其中一个通道进行平均池化操作至 16 ×16,另一通道仍保持 64 ×64。对图像进行平均池化,将多个像素值求和并平均后,可突出背景特征,从而使提取到的特征更加多样。

下一个部分为使用卷积层特征提取模块来对视频中具有的空间信息进行多尺度融合。由于 CTU 编码过程中 CU 的长度为 2 的倍数,因而为了不重叠地提取视频的特征,在特征传递层中卷积操作对应的卷积核(filter)尺寸分别为 4 ×4、2 ×2 以及 2 ×2。将 2 个通道中后 2 个卷积层所提取出的特征,如式(1)所示,输入到后续的全连接层中学习 2 个通道之间的非线性关系。

$$F_{\text{out}} = F_{8\times8\times24} + F_{2\times2\times24} + F_{4\times4\times32} + F_{1\times1\times32} \quad (1)$$

分别经过 2 个全连接层和 softmax 层后输出尺寸为 1 + 2 × 2 + 4 × 4 = 21 大小的划分预测信息 $Info_{\text{split}}$。由于 QP 值的大小对于视频编码质量有着非常大的影响,因此将归一化后的 QP 值作为特征进行了融合。由于 HEVC 标准规定了 52 个量化步长,对应于 52 个 QP(0 ~51),因此将 QP 值通过与 $\frac{1}{51}$ 相乘归一化至 0 ~1 之间。将归一化后的 QP 值与第 1 个全连接层的输出进行拼接,将特征组合到一起,随后进行下一步全连接操作,将 QP 值与特征进行进一步融合。最终得到的 $Info_{\text{split}}$ 将用于判断划分深度为 0 的 64 ×64、划分深度为 1 的 32 ×32 以及划分深度为 2 的 16 ×16 尺寸的 CU 是否需要进一步划分。对于本文所研究的问题来说,仅存在 CU

划分以及不划分 2 种状态，因而得到的预测信息 $Info_{split}$ 最终经过与固定阈值 0.5 进行比较，若 $Info_{split}>0.5$ 则进一步划分，否则不再进行下一深度的划分。若在划分深度为 0 时，网络预测得到 $Info_{split}^{64\times64}<0.5$ 成立，则可以提前终止对是否进一步划分的判断。这就是整个两通道多层特征传递卷积神经网络（two-channel feature transfer convolutional neural network，T-CNN）的结构。

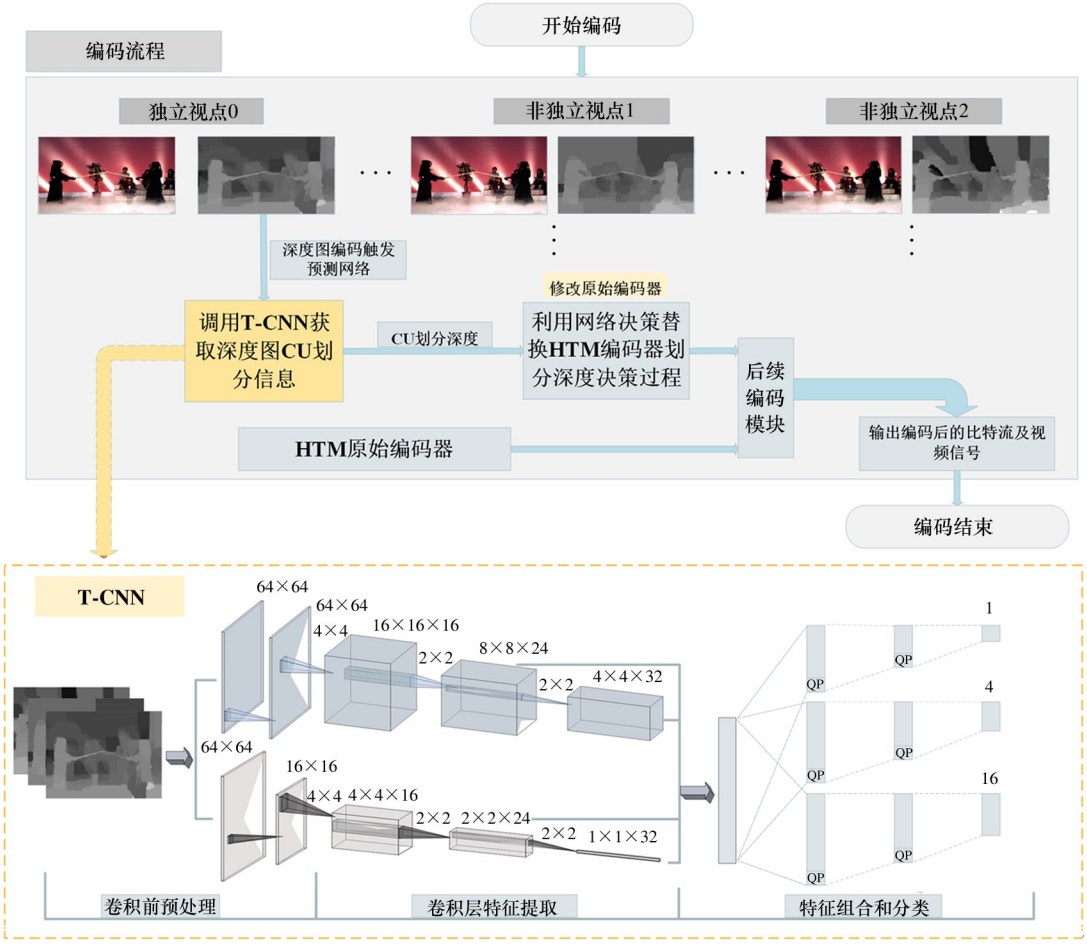


图 6　算法整体流程

此网络将作为划分深度预测模块，在 HTM 标准编码器中替换掉复杂的 CU 划分深度决策过程，加快深度图帧内 CU 划分。具体流程为：开始编码后，在编码到深度图时会触发预测网络，得到预测信息后直接跳过标准的 CU 划分深度决策过程；在 3 个视点中均进行这样的操作，编码器其余部分继续进行后面的编码相关工作，最终输出编码后的比特流以及解码出用于进行质量评估的视频信号，编码结束。

3　实验及分析

3.1　实验设置

实验中使用的是 3D-HEVC 标准测试视频序列：Balloons（1024 × 768）、Kendo（1024 × 768）、Newspaper（1024 × 768）、Poznan _ Hall2（1920 × 1088）、Poznan _ Street（1920 × 1088）以及 Undo _ Dancer（1920 × 1088）。编码时对每个测试序列编码 3 个视点（主视点、依赖视点 1 以及依赖视点 2）。纹理图中的 QP 值以及与其对应的深度图的 QP 值设置为（25，34）、（30，39）、（35，42）和（40，45）。其中，Balloons、Kendo 和 Newspaper 的帧率为 30，Poznan _ Hall2、Poznan _ Street 以及 Undo _ Dancer 的帧率为 25，视频序列的编码帧数为 50 帧。

为了对实验结果有一致的衡量标准，实验均是在配置为 AMD Ryzen 7 4800H、Radeon Graphics 2.90 GHz、64 位 Windows 10 操作系统的计算机上进

行的。训练阶段使用的显卡为 GeForce RTX 2060,实际编码过程中调用模型时仅使用 CPU。为了验证所提出算法的性能,采用全帧内(all intra, AI)编码模式在 3D-HEVC 测试平台 HTM-16.0 上进行测试。编译软件为 Visual Studio 2019,集成开发环境 PyCharm,深度学习库 Tensorflow-GPU 1.13.1。

3.2　评价指标

进行结果分析时,算法的率失真性能使用 BD-rate(bjøntegaard delta bitrate)来进行评价。表 2 中视频 PSNR/视频比特率表示编码纹理视图相对于视频比特率的 BD-rate,视频 PSNR/总比特率表示编码纹理视图相对于总比特率的 BD-rate。

下文中使用 T 代表编码时间,ΔT 代表加入本算法后所节省的编码时间在原始编码时间中的占比,用来表示算法的时间复杂度的下降,其计算公式如式(2)所示。

$$\Delta T = \frac{T_{\text{ori}} - T_{\text{new}}}{T_{\text{ori}}} \times 100\% \tag{2}$$

其中,T_{ori} 为原始编码器 HTM-16.0 的编码时间,T_{new} 为加入本算法后的 HTM 编码时间。

3.3　结果分析

从表 2 可以得出,与原始编码器 HTM-16.0 的编码性能相比,编码视图的平均 BD-rate 损失为 1.4%,率失真性能没有出现明显的下降,在分辨率为 1920×1088 尺寸的视频中,效果要优于低分辨率 1024×768 的视频。图 7 显示了不同 QP 值下原始编码器与加入本文算法后的编码器最终合成的虚拟视点对比图,从主观上可看出并未出现明显失真。

表 2　率失真性能评价

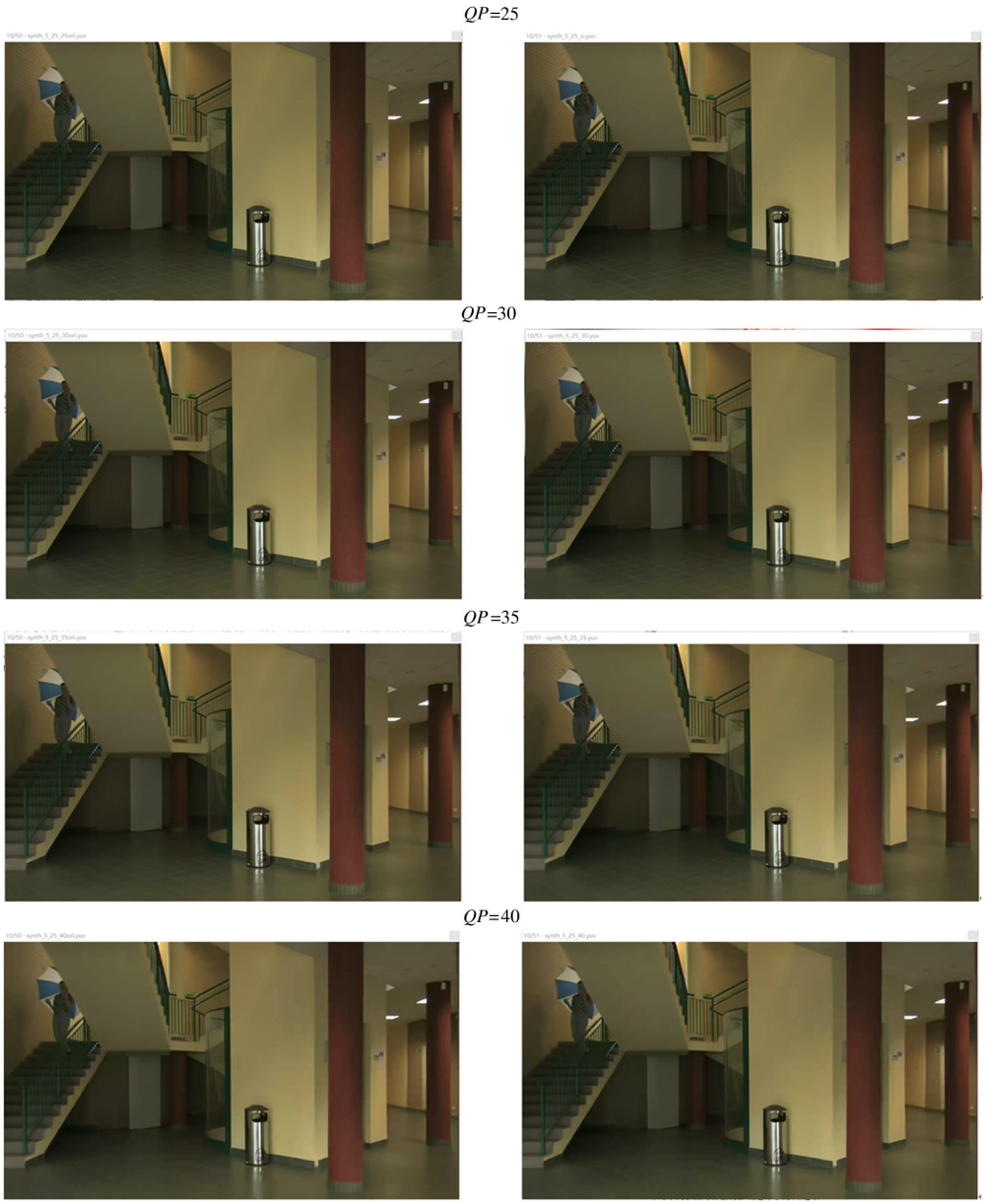
视频序列	视点 0	视点 1	视点 2	视频 PSNR/视频比特率	视频 PSNR/总比特率
Balloons	5.9%	4.7%	5.3%	5.3%	6.5%
Kendo	0.7%	2.3%	0.1%	1.0%	2.6%
Newspaper	1.4%	0.0%	0.0%	0.5%	-0.6%
Poznan _ Hall2	0.1%	0.0%	0.0%	0.0%	-0.4%
Poznan _ Street	0.1%	0.0%	0.0%	0.0%	-0.5%
Undo _ Dancer	1.3%	0.0%	1.8%	1.0%	0.7%
1024×768	2.7%	2.3%	1.8%	2.3%	2.8%
1920×1088	0.5%	0.0%	0.6%	0.4%	-0.2%
平均值	1.6%	1.2%	1.2%	1.3%	1.4%

使用本算法进行单个视点下深度图 CU 划分深度预测所需时间不超过 2.5 s,最多占据编码时间的 0.1%。表 3 将本文算法与其他研究者所提出的算法在编码复杂度降低程度上进行了对比。从表中可以看到,与文献[11,12,15]中的算法相比,加入本文算法后,平均可节省 76.62% 的编码时间,显著降低了编码复杂度。

表 4 列出了实验过程中所用的测试视频所具有的特征,从中可以看到,测试视频的分辨率以及视频所具有的特征种类较为多样。结合表 2 以及表 3 的实验结果,可以得出算法鲁棒性较好,尤其在更高分辨率、具有更多平坦区域的视频中性能出色,可在新的测试数据上进行很好地预测。在对具有少量平坦区域的视频进行预测时,精度仍有提升的空间。

4　结　论

本文通过对 3D-HEVC 标准编码复杂度进行分析,找出编码复杂度过高的深度图帧内 CU 划分过程,针对于这一点建立了深度图帧内 CU 划分深度数据集,并进一步提出了 3D-HEVC 深度图帧内快速编码算法。通过使用两通道多层特征传递卷积神经网络 T-CNN,替代各个视点下深度图帧内 CU 复杂的深度划分过程,可以在保证合成视点质量的同时,显著降低编码复杂度。结果表明,编码时间平均可降低 76%,提升了编码效率。



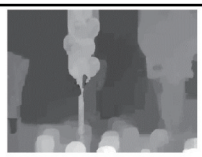
（注：左图为原始编码器，合成视点为 5.25，第 10 帧视频）

图 7　不同 QP 值下原始编码器与加入本文算法后的编码器最终合成的虚拟视点对比图

表 3　编码复杂度比较

视频序列	节省的编码时间占比			
	本文算法	文献[15]	文献[12]	文献[11]
Balloons	79.60%	37.90%	42.50%	46.60%
Kendo	76.00%	36.50%	47.60%	51.30%
Newspaper	78.00%	37.00%	41.60%	36.80%
Poznan _ Hall2	70.20%	35.70%	40.30%	79.10%
Poznan _ Street	78.60%	35.70%	40.90%	61.40%
Undo _ Dancer	77.30%	36.00%	-	69.00%
1024 × 768	77.87%	37.13%	43.90%	44.90%
1920 × 1088	75.37%	35.80%	40.60%	69.83%
平均值	76.62%	36.47%	42.58%	57.37%

表 4　测试视频所具有的特征

视频序列	分辨率	视频图例		视频特征描述
Balloons	1024×768			镜头缓慢平行移动，前景人物有剧烈运动，深度图具有复杂边界信息和部分平坦区域
Newspaper	1024×768			镜头不移动，前景物体轻微运动，背景静止，深度图具有复杂边界信息和部分平坦区域
Kendo	1024×768			镜头平行缓慢移动，前景物体移动明显，深度图具有复杂边界信息和部分平坦区域
Poznan_Street	1920×1088			镜头不移动，前景物体移动明显，背景静止，深度图具有少量集中的边界变化区域和大面积平坦区域
Poznan_Hall2	1920×1088			镜头移动角度大，前景物体运动缓慢，深度图具有少量规则的边界信息和较多平坦区域
Undo_dancer	1920×1088			镜头移动缓慢，前景人物运动较为剧烈，背景包括少量边界和大面积平坦区域

参考文献

[1] WIEGAND T, SULLIVAN G J, BJONTEGAARD G, et al. Overview of the H.264/AVC video coding standard[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2003,13(7):560-576.

[2] SULLIVAN G J, OHM J R, HAN W J, et al. Overview of the high efficiency video coding (HEVC) standard[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2013,22(12):1649-1668.

[3] MERKLE P, SMOLIC A, MULLER K, et al. Efficient prediction structure for multiview video coding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2007,17(11):1461-1473.

[4] KAUFF P, ATZPADIN N, FEHN C, et al. Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability[J]. Signal Processing: Image Communication, 2007, 22(2): 217-234.

[5] Gerhard Tech. 3D-HEVC test model4[C]// Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, 2013:20-26.

[6] KIM I K, MIN J, LEE T, et al. Block partitioning structure in the HEVC standard[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(12):1697-1706.

[7] JCT-VC. HM Software [EB/OL]. (2016-11-05) [2021-09-22]. https://hevc.hhi.fraunhofer.de/svn/svnHEVCSoftware/tags/HM-16.5.

[8] 钟国韵,杨德明,何月顺,等. 基于 Hough 变换的 3D-HEVC 深度图快速帧内预测方法[J]. 东华理工大学学报(自然科学版), 2021,44(5):494-500.

[9] ZHANG R, JIA K, LIU P, et al. Edge-detection based fast intra-mode selection for depth map coding in 3D-

HEVC[C]//2019 IEEE Visual Communications and Image Processing. Sydney, Australia:IEEE, 2020: 1-4.

[10] SALDANHA M, SANCHEZ G, MARCON C, et al. Fast 3D-HEVC depth map encoding using machine learning [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020,30(3):850-861.

[11] FU C H, CHEN H, CHAN Y L, et al. Fast depth intra coding based on decision tree in 3D-HEVC[J]. IEEE Access, 2019,7:173138-173147.

[12] 崔鹏涛,张倩,刘敬怀,等. 基于 FSCD-CNN 的深度图像快速帧内预测模式选择算法[J]. 应用科学学报, 2021,39(3):433-442.

[13] 李雅婷,杨静. 3D-HEVC 深度图帧内预测快速编码算法[J]. 光电子·激光, 2020,31(2):222-228.

[14] XIE S, TU Z. Holistically-nested edge detection[J]. International Journal of Computer Vision, 2015,125(1-3): 3-18.

[15] ZHANG R, JIA K, LIU P, et al. Fast intra-mode decision for depth map coding in 3D-HEVC[J]. Journal of Real-Time Image Processing, 2020,17(5):1637-1646.

Fast intra coding algorithm for 3D-HEVC depth map based on T-CNN

YU Yuan, JIA Kebin

(Faculty of Information Technology, Beijing University of Technology, Beijing 100124)

(Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, Beijing 100124)

(Beijing Laboratory of Advanced Information Networks, Beijing 100124)

Abstract

Depth maps with large flat areas, steep edges, and low texture complexity have been introduced into the 3D-HEVC standard. To solve the problem of high encoding complexity caused by coding unit (CU) rate-distortion optimization of the depth map, a depth map partition dataset is constructed by analyzing the characteristics of the coding process of depth map. And a partition depth prediction algorithm is proposed based on the two-channel feature transfer convolutional neural network (T-CNN). The CU division process of the depth map is replaced by the proposed algorithm under each viewpoint in the original encoder, and the encoding time of the original HTM-16.0 encoder is reduced by about 76% on average with certain loss of rate-distortion performance. It shows that the proposed algorithm significantly improves the coding efficiency.

Key words: 3D-HEVC, depth map, intra-frame coding, convolutional neural network