HEVC 视频编码比赛研究报告

周煜威　　陈昱霏　　蔡沛东　　张子禾

July 2020

1 比赛规则

1.1 比赛目的

- 分析 HEVC 编码标准，利用 SSIM、VMAF、PSNR 等指标，与其他编码标准的最佳实现进行比较。
- 验证各编解码器性能，通过共享测试集数据、编码参数和编码器版本，便于开发人员复现测试结果。

1.2 评判范围

- 客观数据与主观分析
- 编码时间
- 比特率控制
- 速度质量权衡
- 平均结果与不同用例的最优结果

1.3 比赛要求

- 测试集：
  - 50-100 个 HD 视频
  - 10-12 个 4K&HDR 视频

  - 视频的数量会根据参赛编码器的数量增加

- 测试环境：
  - CPU: Intel Socket 1151 Core i7 8700K (Coffee Lake) (3.7Ghz, 6C12T, TDP 95W)
  - Mainboard: ASRock Z370M Pro4
  - RAM: Crucial CT16G4DFD824A 2x16GB (totally 32 GB) DIMM DDR4 2400MHz CL15
  - OS: Windows 10 x64 and Linux (Ubuntu)

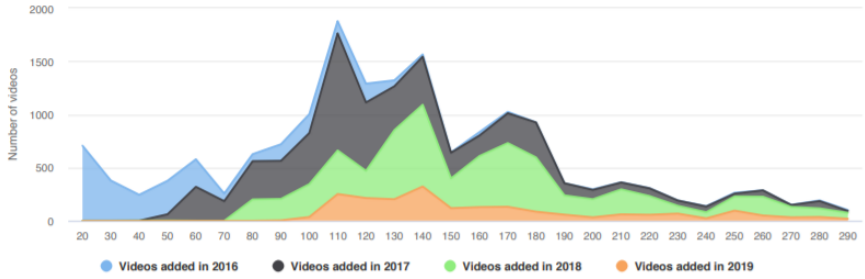
- 编码器要求：
  - 开发人员应提供针对不同速度要求的预设
  - 编解码器应允许以恒定质量模式设置结果流的任意比特率
  - 首选的编解码器界面-控制台编解码器版本（具有批处理支持-必须可以从命令行分配比特率和文件名）
  - 主办方为参与 H.264/AVC 和 H.265/HEVC 标准的编码器提供了参考解码器，所以参赛编码器需兼容参考解码器
  - 允许多线程编码，也不限制 GOP、intra-period、pass，但压缩时间要满足要求

- 编码器预设：
  - 客观比较:
    * Real-time –1080p@30fps
    * Offline –1080p@1fps and SSIM-RD curve better than x264-veryslow
  - 主观比较:
    * Real-time –1080p@30fps
    * Offline –1080p@1fps
  - 4K(UHD) 和 HDR 视频的客观比较:
    * Fast –20fps
    * Universal –1fps

2 数据集分析

2.1 数据来源

主办方自 2016 年开始通过某种选择算法来创建具有代表性的视频序列的集合，在 2019 年从 Vimeo 上抓取了 384946 个视频，从中筛选出至少 50Mbps 的 4K 和 FullHD 视频，最终新增了 145 部 4K 视频和 603 部 FullHD 视频。下图为历年来数据集变化情况：



Year	FullHD videos	FullHD samples	4K videos	4K samples	Total (videos)	**Total (samples)**
2016	3	7	882	2902	885	2909
2017	1996	4638	1544	4561	3540	9299
2018	4342	10330	1946	5503	6288	15833
2019	4945	12402	2091	6016	7036	**18418**

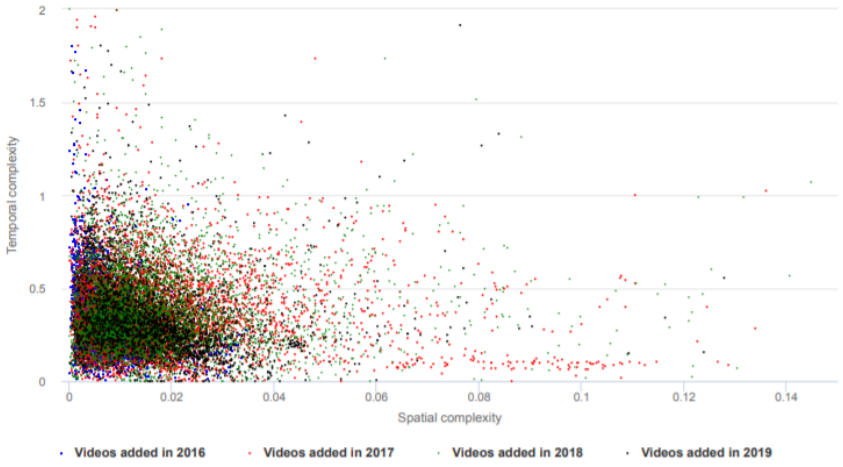
为了避免压缩失真，将 4K 视频调整和裁剪为 FullHD，同时针对场景变化，将所有视频以大致 1000 帧的长度剪切成样本。除去来自 748 个新下载视频的 2585 个样本，数据集也包含之前使用过的 15833 个样本，因此 2019 年的样本数据库总共有 18418 项。

2.2 复杂度

为了评估空间和时间复杂度，使用 x264 采取相同的量化参数对所有样本进行编码，计算每个场景的时间和空间复杂度。

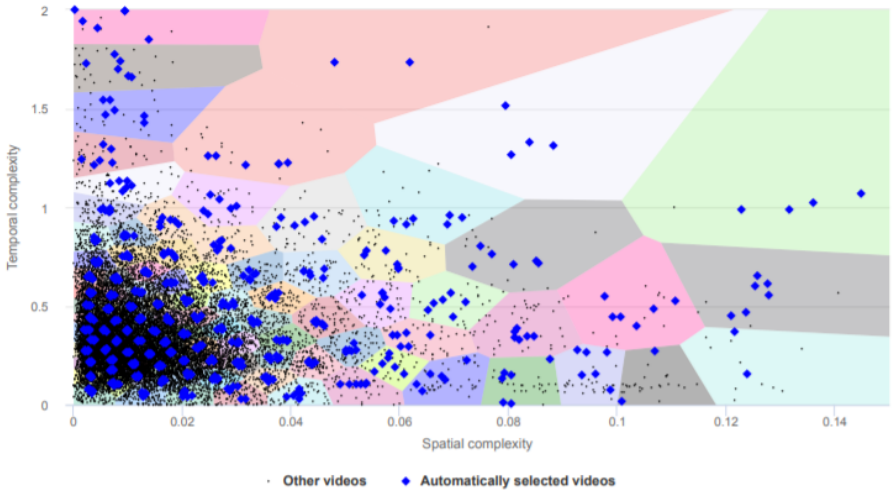
- 空间复杂度：I-frame 的平均大小归一化到样本未压缩帧的大小
- 时间复杂度：P-frame 的平均大小除以 I-frame 的平均大小

此外增加预处理操作来统一视频的色度抽样以避免影响复杂度的评估，即把所有视频转换为 YUV 4:2:0 色度抽样。下图为历年来数据集复杂度分布情况：



2.3 抽样规则

2019 年选取了 100 个样本作为最终的测试集，因此将所有样本分为 100 个集合，从中各随机挑选出 2 到 6 个接近集合中心的候选样本。下图为候选样本分布情况：

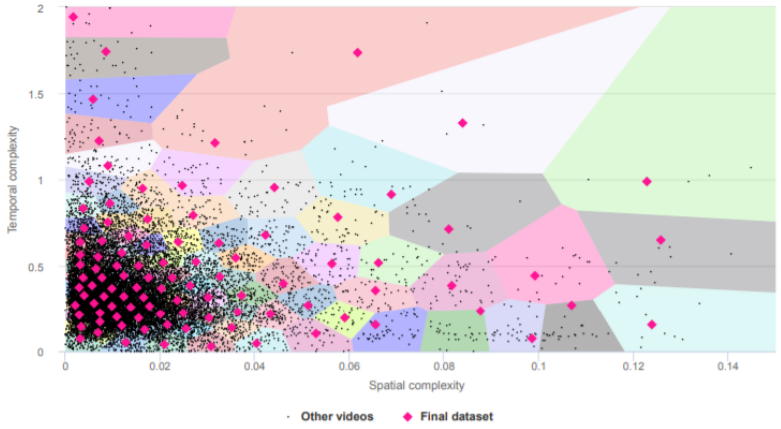


七家参加比赛的公司和两位组织者 (Dmitriy Vatolin 博士和 Dmitriy Kulikov 博士) 以及行业专家 (Jan Ozer) 参与了最终视频集的投票。每个投票者的投票范围为 100 个集群的一个子集，并建议其在每个集群中只选择一个样本。投票者的投票范围是随机选择的，投票范围在不同的投票者间有所重叠并且覆盖了全部的 100 个集群。投票者直至投票结束前都可更改投票，最终每个集群中得分最高的样本共同组成测试集。下图为投票者情况：

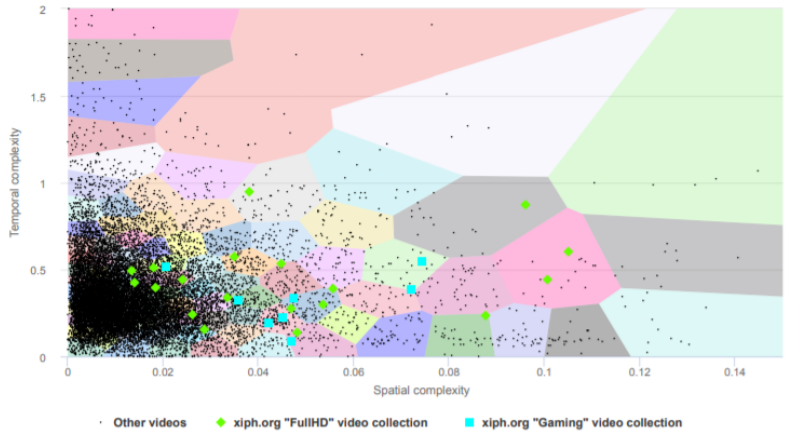
Voter	Number of clusters to vote	Number of received votes	Vote weight
Dr. D. Vatolin	100	100	1
Dr. D. Kulikov	100	100	1
Jan Ozer	70	70	2
Participant #1	25	25	1
Participant #2	25	25	1
Participant #3	25	25	1
Participant #4	25	25	1
Participant #5	25	15	1
Participant #6	25	8	1
Participant #7	25	7	1

2.4 抽样结果

最终的数据集由 100 个视频组成，其中 8 个来自旧数据集，92 个来自 Vimeo 和 xiph.org；平均比特率为 218.9 Mbps，中位数为 143.2 Mbps。下图为数据集分布情况：



通过比较 xiph.org 和 Vimeo 的数据，可以看出 xiph.org 上的大部分视频具有编解码器在日常生活中很少遇到的较高的空间以及时间复杂性。下图为比较结果分布情况：



3 评价指标

3.1 概述

多数视频编码器都支持指定编码的码率，编码后的视频质量与码率一般正相关，因此可以通过码率-质量图来表示其压缩性能。通过曲线图的对比，可以比较不同编码器在相同质量下的码率，或相同码率下的视频质量。

3.2 码率

平均码率即单位时间的视频编码后的大小，单位为 Mbps。同样的视频编码后码率越低，则压缩率越高。

3.3 质量指标

3.3.1 PSNR(Peak Signal to Noise Ratio)

PSNR 通过峰值信噪比反映编码后的图像与原图像之间的差别，即压缩后的图像质量。其中噪声量用 MSE(Mean Squared Error) 表示，并用使

用信号最大可能值的平方除以该噪声，得到归一化的峰值信噪比:

$$\mathrm{MSE} = \frac{1}{mn}\sum_{i=1,j=1}^{m,n}(Y_{i,j} - X_{i,j})^2$$

$$\mathrm{PSNR} = 10 \cdot \log_{10}(\frac{\mathrm{MaxErr}^2}{\mathrm{MSE}})$$

其中 MaxErr 表示 $|Y - X|$ 的最大可能值。当计算一段视频的 PSNR 时，将所有帧视为一整张图像计算 MSE。

PSNR 是比较通用的信号质量评价标准，可以用于 R,G,B,Y,U,V,L 空间。PSNR 度量方法计算简单、快速，但有时并不能很好地反映人类对图像/视频的感知。

该指标越高图像质量越好。

3.3.2 SSIM(Structural Similarity)

图像上的卷积运算:

$$\langle U, W\rangle = \sum_{i=-R,j=-R}^{R,R} U(x+i, y+i)W_{i+R,j+R}, \text{ where } R = \frac{N}{2}$$

其中窗函数 W 为 $W_{i,j} = \frac{1}{(N+1)^2}$ 或者高斯窗函数 ($\sigma = 1.5, N = 10$)

$$\begin{aligned}
\mu_x &= \langle X, W\rangle \\
\mu_y &= \langle Y, W\rangle \\
\sigma_x &= \left\langle (X-\mu_x)^2, W\right\rangle \\
\sigma_y &= \left\langle (Y-\mu_y)^2, W\right\rangle \\
\sigma_{xy} &= \langle (X-\mu_x)(Y-\mu_y), W\rangle
\end{aligned}$$

每个像素的 SSIM 定义如下:

$$\mathrm{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x + \sigma_y + C_2)}$$

其中 $C_1 = 0.01^2, C_2 = 0.03^2$。总 SSIM 为所有像素 SSIM 的平均值。
SSIM 还可以从亮度、对比度、结构三方面度量图像相似性:

$$l(X,Y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$
$$c(X,Y) = \frac{2\sqrt{\sigma_x\sigma_y} + C_2}{\sigma_x + \sigma_y + C_2}$$
$$s(X,Y) = \frac{2\sigma_{xy} + C_3}{\sqrt{\sigma_x\sigma_y} + C_3}$$

该指标越高图像质量越好。

3.3.3 VMAF(Video Multimethod Assessment Fusion)

VMAF 是由 Netflix 与南加州大学合作开发的视频质量指标。VMAF 通过将多种基本的质量指标结合在一起预测主观质量，面对不同特征的源内容、失真类型，以及扭曲程度，每个基本指标各有优劣。通过收集主观质量数据，并使用 SVM 将基本指标“融合”为一个最终指标，可以保留每个基本指标的优势，以得出更精确的最终分数。

VMAF 的结果非绝对指标，不同视频、不同分辨率的 VMAF 得分不能直接比较。

该指标越高图像质量越好。

在之后的任务中，主要采用 SSIM 作为视频质量的指标。

4 比赛结果

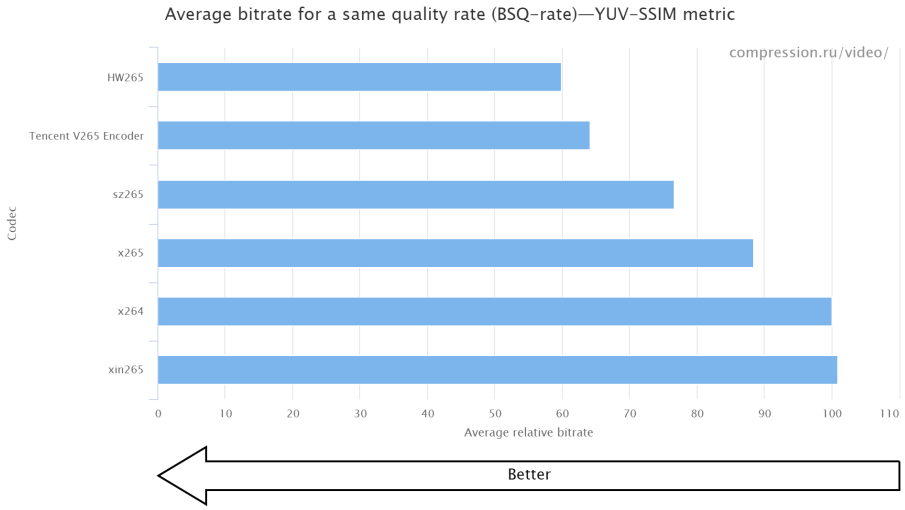
比赛从多个角度评价了各编码器性能，其评判标准在本次报告的第一部分中已经有所提及。

由于本次报告仅对编码器性能进行总体上的概述，因此选取了其中两个评判标准，即客观数据和主观打分，从而对各编码器在不同维度、不同样例和不同条件下的表现进行简述。

4.1 客观数据 (YUV-SSIM)

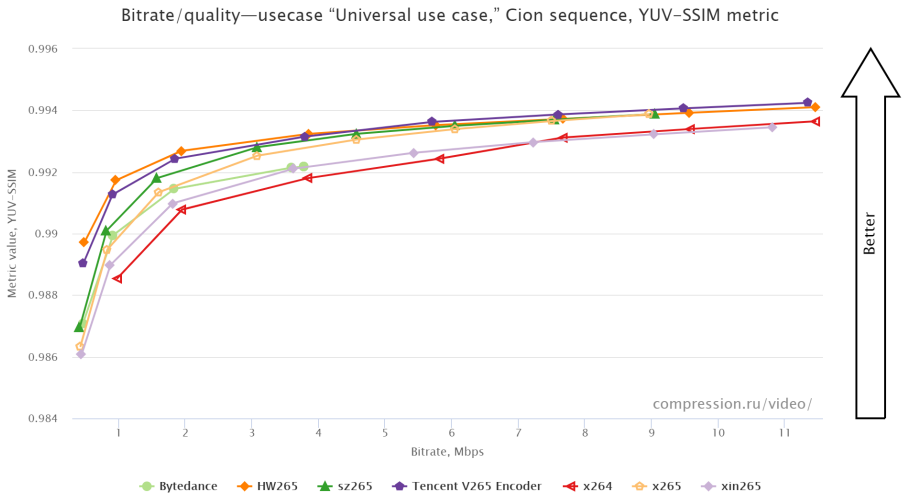
4.1.1 质量相同条件下的平均码率图

由下图，将 YUV-SSIM 作为质量指标时，平均码率最小的编码器分别为：HW265, Tencent V265 Encoder 与 sz265。



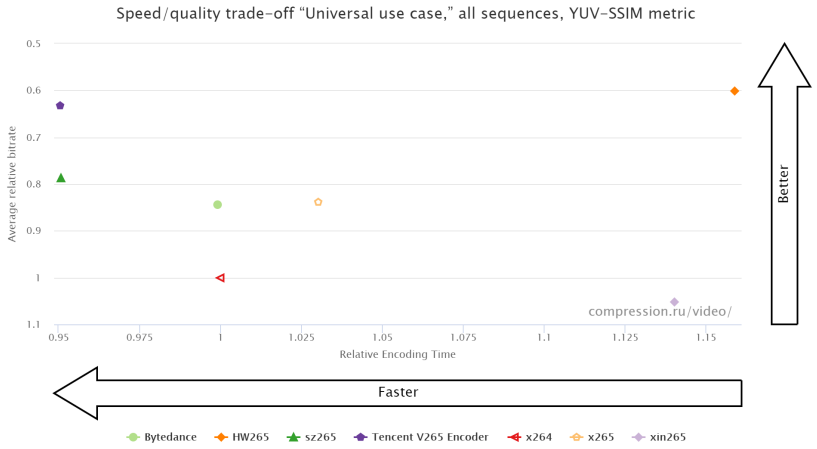
4.1.2 通用编码的码率-质量图

由下图，比较图中各编码器编码质量随码率的变化曲线，可以观察到，左上角折线所代表的编码器要优于右下角折线所表示的编码器。也就是说，相同码率时左上角的质量更好，相同质量时左上角的码率更小。因此，更优的编码器为：HW265 与 Tencent V265 Encoder。



4.1.3 通用编码的速率-质量图

由下图，图中各点的位置越靠上，编码质量越优；位置越靠左，编码速率越快。其中，Tencent V265 Encoder 在图像的最左上方，因此更优。



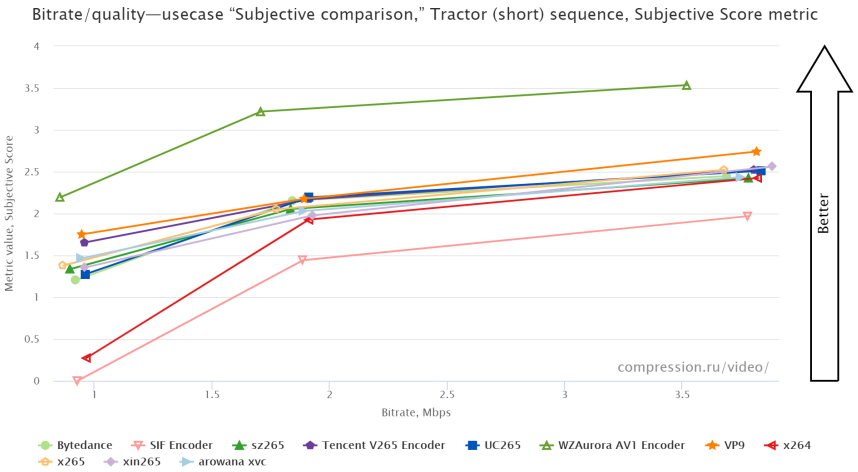
4.2 主观分析

4.2.1 质量相同条件下的平均码率图

由下图，将主观打分作为质量指标时，平均码率最小的编码器为 WZAurora AV1 Encoder。

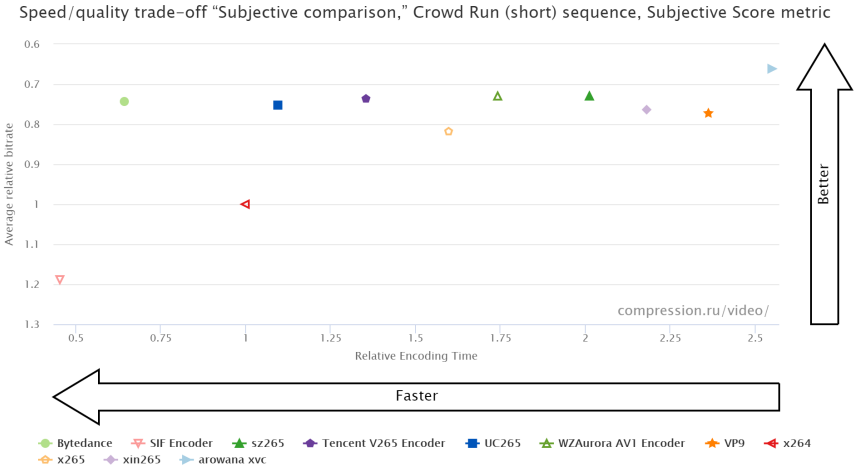
4.2.2 通用编码的码率-质量图

由下图，左上角折线所代表的编码器要优于右下角折线所表示的编码器，相同码率时左上角的主观评分更高，相同评分时左上角的码率更小。因此，更优的编码器为 ByteDance V265 Endcoder。



4.2.3 通用编码的速率-质量图

由下图，图中各点的位置越靠上，主观评分越高；位置越靠左，编码速率越快。其中，ByteDance V265 Encoder 在图像的最左上方，因此更优。



5 参考

- HEVC/AV1 Video Codecs Comparison 2019
- MSU Annual Video Codecs Comparison 2019: Call for codecs
- MSU Main report - Free Vision
- A Subjective Study for the Design of Multi-resolution ABR Video Streams with the VP9 Codec
- MSU Quality Measurement Tool: Metrics information
- 图像质量评价指标 PSNR、SSIM、MSSIM 介绍