【本文献信息】王红胜，徐超，曹凯.4K超高清图像质量客观评价工程实践[J].广播与电视技术，2021，Vol.48(10).

4K超高清图像质量客观评价工程实践

王红胜，徐超，曹凯

（国家广播电视总局广播电视科学研究院，北京 100866）

【摘 要】超高清视频技术得到了越来越广泛的应用，如何对其质量进行评价也就显得尤为重要了。本文在陈述当前图像评价的常见方式的基础上，选取了当前主流的全参考评价的PSNR、SSIM和VMAF三种评价方式，对经过AVS2和HEVC编解码的图像质量进行了分析比对，并给出了相关结论和建议。

【关键词】超高清，图像质量，客观评价

【中图分类号】TP306 **【文献标识码】**A **【DOI编码】**10.16171/j.cnki.rtbe.20210010002

Engineering Practice of Objective 4K UHD Image Quality Assessment

Wang Hongsheng, Xu Chao, Cao Kai

(Academy of Broadcasting Science, NRTA, Beijing 100866, China)

Abstract UHD video technology has been more and more widely used, and how to evaluate its quality becomes particularly important. Based on the description of current common methods of image evaluation, this paper selects current three mainstream full-reference evaluation methods PSNR, SSIM and VMAF, analyzes and compares image quality encoded and decoded by AVS2 and HEVC, and gives relevant conclusions and suggestions.

Keywords Ultra High-Definition, Image Quality, Objective Assessment

0 引言

随着超高清电视技术的发展，4K超高清频道的不断播出，超高清概念已经深入人心。在传输端基于AVS2编码和HEVC编码方式的图像压缩编码方式也已经得到了广泛的应用。视频图像质量评价作为视频图像制作、传输、播出和接收的重要方面，如何发挥更为有效的作用，已经变得尤为重要了。

1 基本概念

1.1 4k超高清

K可以用来表示色温高低的测量单位开尔文，也可以表示为功率的单位千瓦，但是在数字图像中，1K代表1024像素或感光点。4K从技术上来讲其分辨率应该是4096×2160，但是我们通常意义上认为分辨率是3840×2160。超高清（Ultra High-Definition）是ITU批准的4K分辨率（3840×2160）的正式名称。

1.2 AVS2和HEVC

在最新的视频压缩编码标准中，我国和其他国家采用了不同的编码标准。我国采用的是具有自主知识产权的AVS2编码标准，国际标准是HEVC（高效视频编码），俗称H.265，其最终版本已被ITU-T视频编码专家组接纳批准成为国际标准，通常认为是AVC（H.264）的继承者。

1.3 图像质量评价

图像视频经过一系列的压缩编码传输解码等过程，最终的目的是将蕴含的大量有价值的视觉信息呈现给终端的用户。所以最终呈现的图形视频质量好坏直接影响着终端用户的直观感受。这个结果受诸如图像质量、显示器的等级等多方面因素的影响，而其中图像质量本身的好坏起着决定性的作用。

图像质量评价一般分为客观评价和主观评价。

主观评价国内目前主要采用GY/T 134-1998《数字电视图像质量主观评价方法》标准，对应的国际标准是ITU-R BT.500系列标准，目前最新的是BT.500-13“Methodology for

the subjective assessment of the quality of television pictures”。其一般采用双刺激连续质量标度法或者双刺激损伤标度法，两者方法都需要对观测者连续给出原始视频图像和经过被测状态的视频图像，由观测者给出主观感知分值[1]。

客观评价其实是根据人眼的主观视觉系统建立相应的视觉感知数学模型，并通过不同的数学公式或统计算法计算图像视频的质量。根据评价时是否需要参考图像，客观评价又可以分为全参考、半参考和无参考等三类评价方法[2]。全参考评价需要无失真的原始图像和所测图像进行比对，得到评价结果；半参考方法需要将所测图像的某些特征和无失真的原始图像利用小波变换系统的概率分布或多尺度几何分析等进行比对；无参考评价则无需无失真的原始图像，根据图像自身的特征来估计图像的质量，需要大量的图像质量评价数据库等。目前工程学的应用上主要采取全参考的图像质量评价方法，本文的研究结果也是基于全参考的图像客观评价方法。

2 全参考图像质量客观评价

全参考图像质量客观评价的算法有很多种，每种算法都是基于一定的数据集和相关公式，如何对其进行评价需要有一定的衡量标准。

2.1 方法概述

全参考图像质量客观评价目前应用比较多的是峰值信噪比（PSNR）和均方误差（MSE），两者都是基于像数统计特性的；还有比较经典的差异评价主观得分（DMOS），表示的是被测图像和原始图像之间的差值，由于其能够更好的反映图像的相对失真程度，被图像质量评价数据库选做主观评价结果；后来为了更接近人类视觉系统（HVS）的评价结果，出现了结构相似性（SSIM）评价方法。还有NQM、UQI、MS-SSIM、IFC、VIF、IW-SSIM、RFSIM、FSIM等多类评价准则。而VMAF是美国Netflix公司开发的一套视频质量评价体系，将人类视觉模型与机器学习相结合，使其面对不同特征的源内容、失真类型，以及扭曲程度，对于各有优劣的各项基本指标，通过使用机器学习算法将基本指标“融合”为一个最终指标。每个视频图像评价算法都会基于一定的视频图像模型库，评价图像客观评价算法的好坏主要是在具有不同失真的大数据集上观察者的主观评分和算法评分的相关度，如果相关度越高，该评价算法的性能越好，反之亦然。

2.2 图像模型数据库和优劣模型

全参考、半参考和无参考的图像质量评价过程中，需要加入各种数据库模型，这样就需要各种类型的图像模型数据库。数据库由大量的经过人工评价的图像和相应的分值，我们这里介绍几种典型的数据库模型。TID2013、TID2008、CSIQ、LIVE、IVC、MICT、WIQ和A57，这些数据库模型对于图像视频质量的研究和评价提供了基础数据，通常以算法评价和主观评价来对其误差和相关性进行评价。一般情况下，DMOS值越小，说明图像质量越好，相关性越强，反之亦然。一般情况下，广泛采用的技术指标有以下几个，分别是RMSE（均方根误差）、LCC（线性相关系数）、SROCC（Spearman秩相关系数）和KROCC（Kendall秩相关系数）等[2]。

2.3 三种算法表达

本文中，我们将使用PSNR、SSIM和VMAF三种主流的评价方式。

1. PSNR

PSNR（Peak Signal to Noise Ratio），即峰值信噪比，是一种比较简单的全参考图像质量评价指标，也是使用最广泛的一种视频图像客观评价指标。可以得出被测图像相对原始图像的相似度，其计算公式如下：

$$PSNR = 10\log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left[I(i,j) - K(i,j)\right]^2$$

$I(i,j)$ -original pixels

$K(i,j)$ -pixels distorted by lossy compression

其取值一般介于30～40之间，取值越高，则代表图像质量越好。

2. SSIM

SSIM（Structural Similarity Index Measurement），即结构相似度测量，也是一种全参考图像质量评价指标，其计算公式如下：

$$SSIM(x,y) = \left[l(x,y)^{\alpha}\cdot c(x,y)^{\beta}\cdot s(x,y)^{\gamma}\right]$$

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \qquad c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}$$

$$s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \qquad c_3 = c_2/2$$

该算法实际应用时计算像素所在窗口区域的SSIM值，再求其均值作为整幅图像的SSIM值。其通常的取值区间是0.9～1，取值越高，则代表图像质量越好。

3. VMAF

不同于传统的PSNR和SSIM评价模型，视频多方法评估融合VMAF（Video Multimethod Assessment Fusion）是由Netflix

提出的一种基于人眼主观质量来训练模型并考虑了时间相关性的视频质量评价模，是一种更多的参考了主观意识的客观评价标准。该方法主要是基于SVM的nuSvr算法，通过事先训练好的模型，对每种视频特征以不同的权重，对每一帧画面生成一个评分，最终以均值算法得出该视频的最终评分。

通过衡量视觉信息保真度、细节丢失指标和运动量三个方面的信息得到主观质量模拟打分。VMAF目前有HDTV、Phone和4K超分辨率等多种开源评价模型可供使用。我们此次使用的是4K超分辨率评价模型。

其取值范围可以是0～100，一般75～85之间的取值我们就认为图像质量就比较好了。相较于PSNR和SSIM评价的快速运行，VMAF运行时间大概是PSNR的十倍左右。

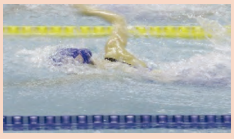
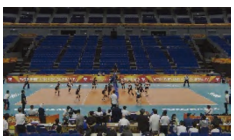
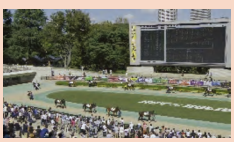
3 AVS2和HEVC编码测试结果比对

为了使得节省传输的带宽，我们通常会对包含大量数据信息的图像视频进行编码压缩处理，然后在接收端解码输出，而这个过程由于编码程度和方式的不同，就会带来不同的图像损伤，所传输的图像视频会引入负面因子，如何评判这些负面因子的影响，就需要通过上述质量评价的方式来进行。

目前主流的4K编码方式主要是AVS2和HEVC，我们选取了六个经典的超高清原始图像序列作为原始参考，对经过不同编解码方式输出的图像选择了PSNR、SSIM、VMAF三种客观评价方法进行评价。

本次测试采用6个测试图像序列，图像序列格式为3840×2160/50/P，色域标准为ITU-R BT.2020，具体图像序列描述如表1所示。

表1 测试序列信息列表

序号	测试序列名称	序列描述
1	 Fireworks	该序列是可以目视识别的燃烧的单色烟花，以夜视图的一个大厦为背景。高亮度烟花在接近黑色的暗区背景下缓缓上升，当烟花爆炸时达到最大亮度，光随着烟花变化而变暗，形成高亮度渐变和高亮度压缩特性。在黑暗的一面，烟花周围确认到部分黑色的漂浮物为光晕。主要考察编码系统的清晰度、宽色域和亮度细节处理能力
2	 Drama	该序列为客厅就坐女演员的起立场景。它从女演员就座的场景开始，当女人站起来时，平移相机的摄像头，放在桌上的花瓶中的花朵部分具有多种颜色，女性服装，背景中风帆起皱的部分以及家具的木纹纹理部分的细度非常重要。主要考察编码系统的色彩还原及亮度细节处理能力
3	 Sunset	序列内容是背景为夕阳的海滩景色。这是不涉及相机操作的固定拍摄。滑浪风帆作为移动的物体穿过屏幕的中心，在海面上，夕阳反射的区域具有较高的亮度，其他区域具有较低的亮度。相对较大的波表面涉及复杂的运动造成屏幕底部的起伏，从屏幕中心附近向左穿过的滑浪风帆涉及连续的渐进运动。主要考察编码系统的时间变化特性、高饱和色彩还原的处理能力
4	 Swim race	序列内容是一个游泳比赛视频，通过从侧面跟踪运动员如何游泳来构成场景。就亮度和颜色而言，它相对单调，画面主要跟随运动员的平移操作相对较快，并且运动员的运动幅度也较大。主要考察编码系统的处理的准确性和鲁棒性、快速运动的处理能力
5	 Volleyball	该序列是一场排球比赛的视频，整个球场都被捕获在屏幕上。运动员运动复杂，运动员之间经常发生咬合，并且球运动的速度和方向频繁变化，因此适合进行运动适应性处理；此外，前景中的球场是照亮的，而后面的观众席是黑暗的。主要考察编码系统的清晰度、高饱和度和色彩还原的准确性和运动预测补偿性能
6	 Paddock	序列内容是围场在赛道上的固定图像。后半部分是一张松散的镜头，其中包括有阴影的观众座位和围场。 背面的草木茂盛，细节感十足，可以评估纹理的表现力和视频编码所导致的失真。 另外，大显示屏上显示的精细字符适合评估分辨率。 另外，它是晴天的室外图像，对比度非常高的图像。主要考察编码系统的清晰度、高对比度和高动态范围细节等的处理能力

3.1 AVS2编码测试结果

根据上述内容，针对PQR/VMAF/SSIM三种不同的评价方式绘制了在不同码率编码情况下的变化曲线，如图1～图3所示。

图1～图3分别展示了在Fireworks、Drama、Sunset、Swim race、Volleyball和Paddock六个不同序列图像在不同编码码率情况下的图像评价变化曲线。变化曲线的横轴是编码码率，变化曲线的纵轴是相对应的取值。AVS2（1）、AVS2（2）、AVS2（3）表示三种不同的模式，分别是Ultrafast、Media和Dmos，是压缩效率和运算时间平衡的一个预设值，默认为Media。从变化曲线上我们可以看出。

1. 随着编码码率的提高，PQR/VMAF/SSIM三种不同的评价方式的取值不断提升，这也符合编码码率越高，图像损伤越小的基本规律；

2. Ultrafast、Media模式下的值比较接近，Dmos模式下的值和Ultrafast、Media模式下的值相比差异比较明显，比前两种模式下的值要高很多；

3. 从曲线变化的趋势上看，PQR/SSIM的取值变化更趋向于一致，而VMAF评价方式对于Drama、Sunset这种静态变化的图像序列评价表现和另外两种稍微有所差异。

3.2 HEVC编码测试结果

图4～图6分别展示了在Fireworks、Drama、Sunset、Swim race、Volleyball和Paddock六个不同序列图像在不同编

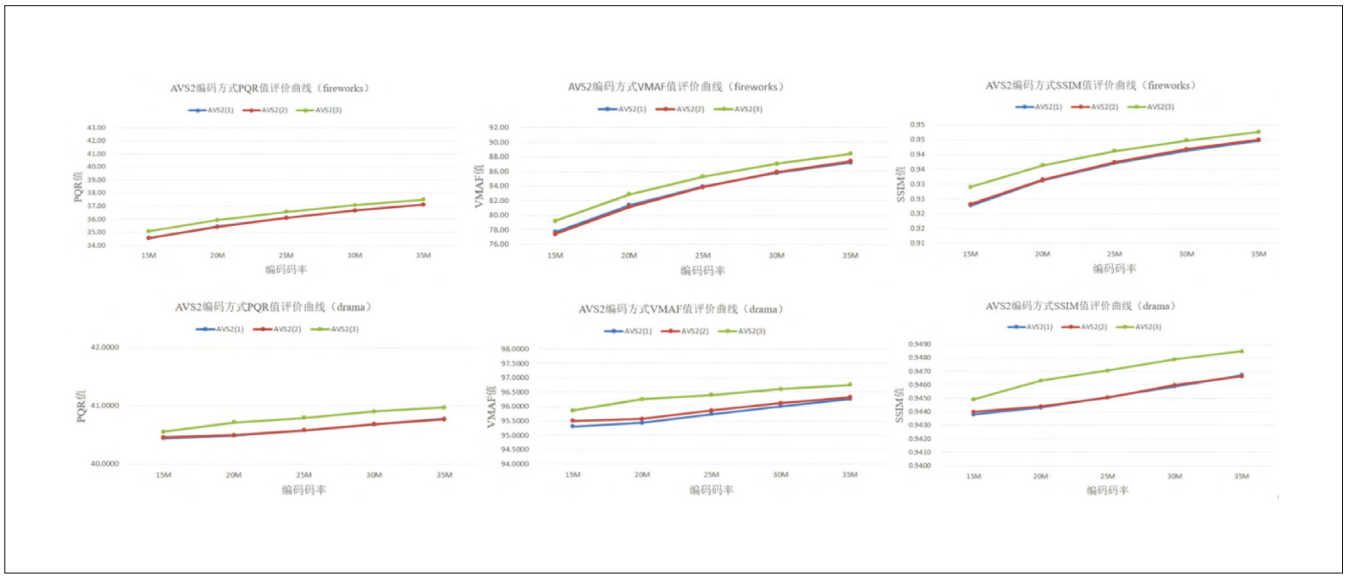


图1 AVS2编码方式PQR/VMAF/SSIM值评价曲线（Fireworks/Drama）

图2 AVS2编码方式PQR/VMAF/SSIM值评价曲线（Sunset/Swim race）

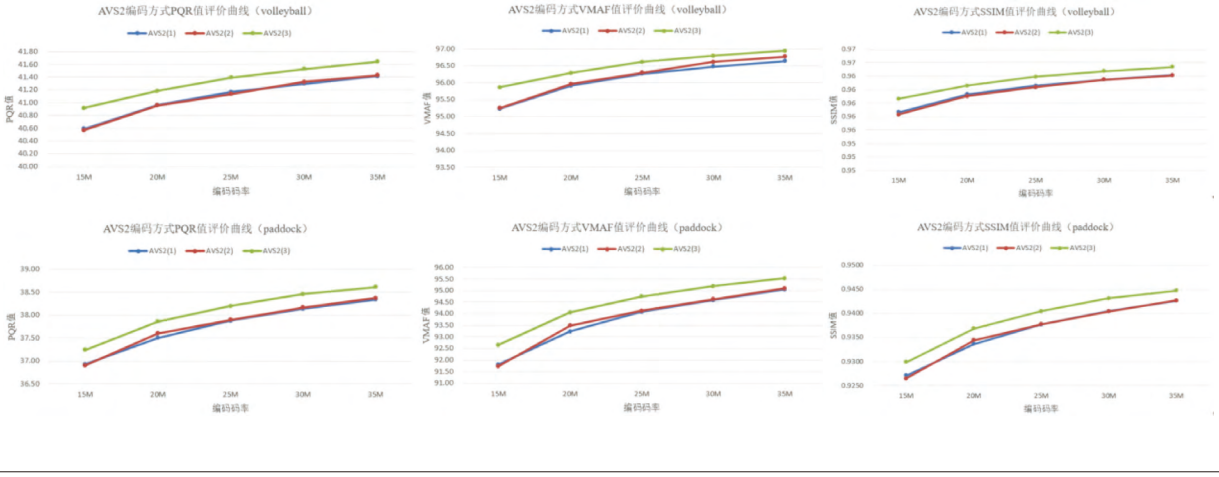


图3 AVS2编码方式PQR/VMAF/SSIM值评价曲线（Volleyball/Paddock）

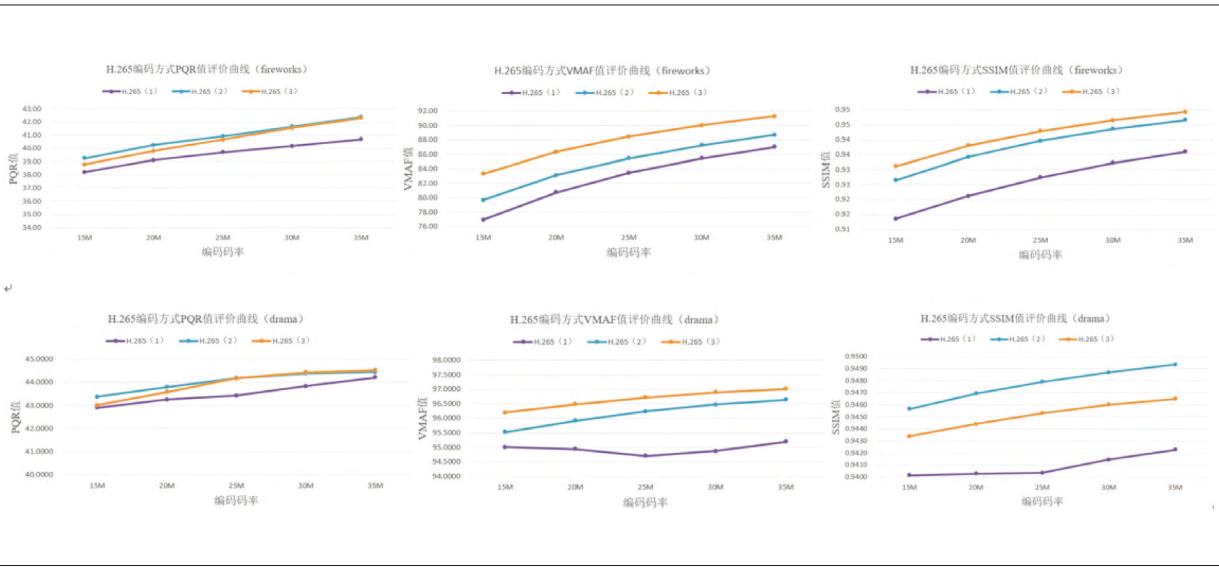


图4 编码方式PQR/VMAF/SSIM值评价曲线（Fireworks/Drama）

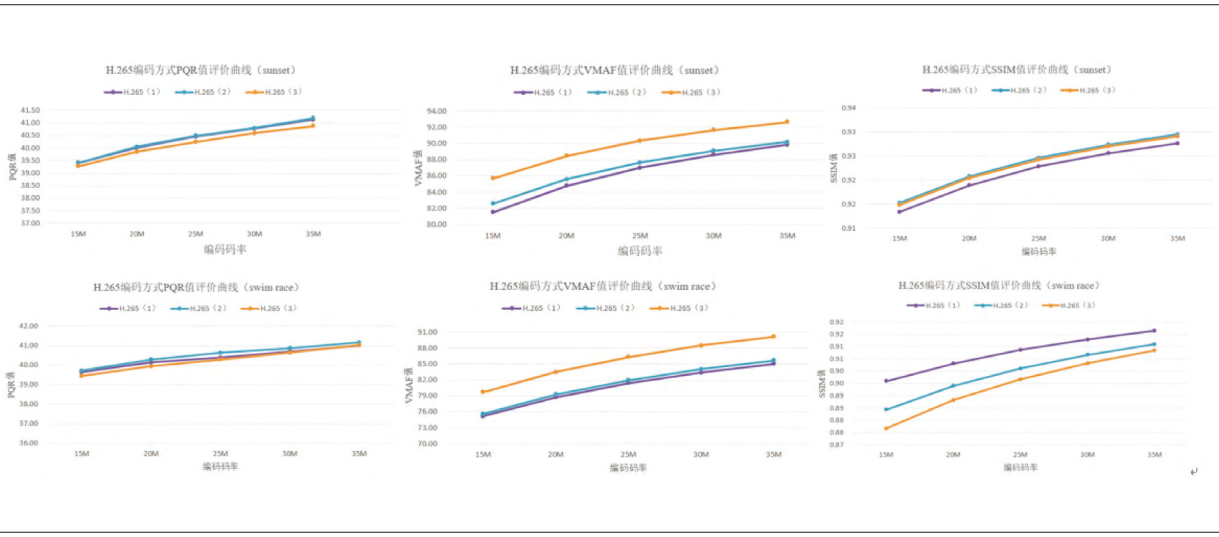


图5 编码方式PQR/VMAF/SSIM值评价曲线（Sunset/Swim race）

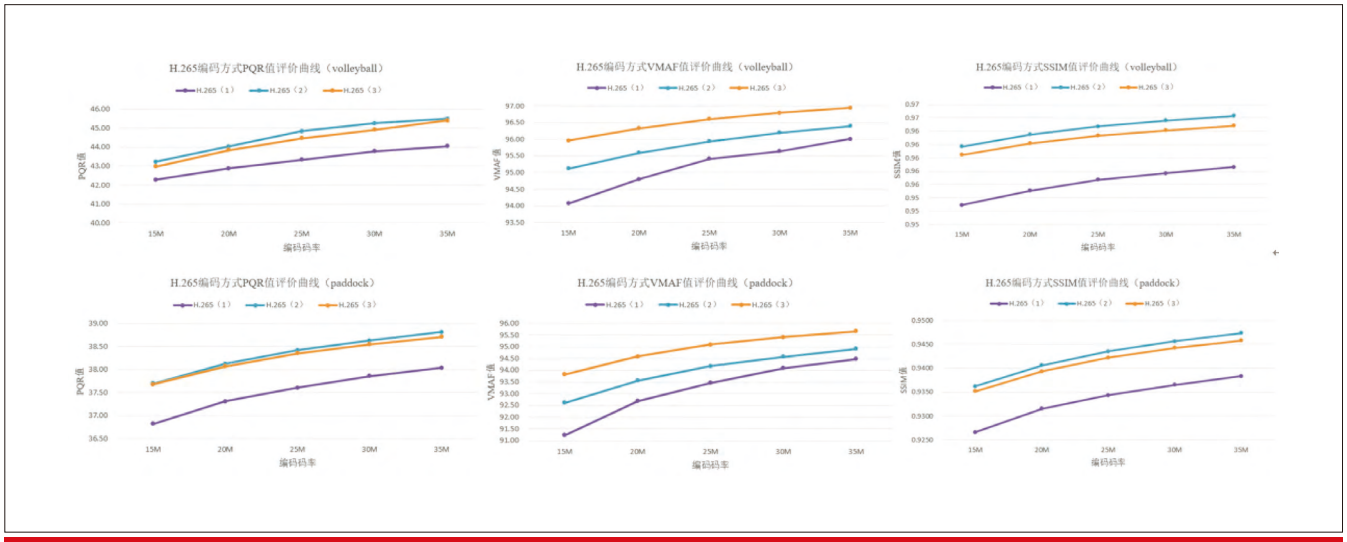


图6　编码方式PQR/VMAF/SSIM值评价曲线（Volleyball/Paddock）

码码率情况下的图像评价变化曲线。变化曲线的横轴是编码码率，变化曲线的纵轴是相对应的取值。H.265（1）、H.265（2）、H.265（3）表示三种不同的模式，分别是Ultrafast、Media和Veryslow。

从变化曲线上我们可以看出。

1. 随着编码码率的提高，PQR/VMAF/SSIM三种不同的评价方式的取值不断提升，这也符合编码码率越高，图像损伤越小的基本规律；

2. Ultrafast、Media和Veryslow三种模式下的取值有的比较接近，有的比较离散，并没有体现出明显的变化趋势；

3. 从曲线变化的趋势上看，VMAF/SSIM的取值变化相对更趋向于一致，而PQR评价方式和另外两种有所差异；

4. 对于Swim race和Paddock这两个对于细节要求较高和变化趋势比较大的图像序列来说，Ultrafast、Media的SSIM取值反而出现了比Veryslow这种模式要好的情况，和PQR/VMAF的取值变化趋势相悖。

3.3 结果比对

在上述的结果比对中，我们可以看出对于三种常见的图像质量客观评价方式，AVS2编码适用于三种评价方式中的任何一种，而H.265编码方式变化趋势规律不太一致，尤其是在SSIM评价方式下，不能正确的反映H.265编码方式的优劣，所以相比较而言，PQR/VMAF评价方式更适合于H.265编码。

4 结论

本文通过主流的PSNR、SSIM、VMAF三种图像客观评价方法对在不同编码方式下的4K超高清图像进行了客观评价和分析，不同的评价方式有各自的优劣点，这就要求我们在实际使用过程中择优进行选择。后期我们将针对这些4K超高清图像按照标准要求进行主观评价，以验证其相关的符合性。RTBE

参考文献：

[1] 王红胜, 徐超, 张为冬. 基于云平台的广电融媒体制播平台测试方法探讨[J]. 广播电视信息, 2020.05, Vol. 29.

[2] 王志明. 无参考图像质量评价综述[J]. 自动化学报, 2015, Vol.41.

[3] 丰明坤, 赵生妹, 邢超. 一种基于视觉特性的PSNR图像评价方法[J]. 南京邮电大学学报(自然科学版). 2015, Vol. 35. No.4.

[4] 杨璐, 王辉, 魏敏. 基于机器学习的无参考图像质量评价综述[J]. 计算机工程及应用, 2018, 54(19).

第一作者简介：

王红胜，男，1986年10月，毕业于中国传媒大学，研究生学历，硕士学位，国家广播电视总局广播电视科学研究院工程师。主要从事广播电视电视中心、网络视听、软件和信息安全等方面的技术研究与测试。