视频质量主观评估的方法和规范

王伟（wangwei1237@gmail.com）

目录

1. 视频质量主观评估体系存在的价值

✦人类视觉系统的特性为数字视频压缩技术提供了可行性。

✦视频质量评估研究的是**视频压缩技术**作用于**人类视觉系统**上的可感知的差异。

✦主观评估视频质量的过程中的变量非常多，如何保证主观评估结果的客观性？有没有对应的理论体系支撑？评估的方法业界同行是否认可？

✦和其他领域不同，视频质量主观评估的**理论体系**是视频质量评估工作的基石。

2. ITU 提出的视频质量主观评估标准

ITU：国际电信联盟，制定全球电信标准，促进全球电信发展。

- ITU-R：无线电通信部门
- ITU-T：电信标准化部门
- ITU-D：电信发展部门

H26X系列的视频编解码标准就是 ITU-T 部门组织并发布。

ITU 提供了视频质量主观评估的标准，其标准也获得了业界的广泛认可。

- 网飞的 VMAF 算法所依赖的数据是基于 ITU 的主观评估标准来获取的。
- 头条的对应视频质量算法所依赖的数据也是基于 ITU 的主观评估标准来获取的。

2. ITU 提出的视频质量主观评估标准——ITU-R

ITU-R BT. 500 建议书（Methodologies for the subjective assessment of the quality of television images）给出了电视图像质量的主观评价方法。

在 BT. 500 中，从以下几个方面给出了主观评估的方案：

- 观看环境的建设
  - 实验室环境
  - 家庭环境
- 评估者的选择
- 评估的方法：控制图像如何展现
- 打分的方法：控制评估者如何给出自己的主观感受
- 打分置信区间的判断

ITU-R BT. 2095 建议书（Subjective assessment of video quality using Expert Viewing Protocol）给出视频质量的专家主观评价方法。

2. ITU 提出的视频质量主观评估标准——ITU-T

互联网下的多媒体视频质量评估的相关标准，大部分集中在 ITU-T 的 P系列。

- P 系列：电话传输质量、电话装置和本地线路网络质量的客观和主观评定方法
- P. 900：多媒体业务的视听质量

目前 P. 900 系列标准包含了 P. 910 - P. 931 的 13 个标准，用于评估各种视频质量的主观评估方法。

- 视频质量的主观评估标准：P. 910，P. 911，P. 913，P. 930
- 视频 QoE 的主观评估标准：P. 917，P.918
- 其他类型的视频质量主观评估标准：P. 912，P.914-P.916，P.919，P.920，P.931

2. ITU 提出的视频质量主观评估标准——ITU-T P. 900

视频质量主观评估标准

- P. 910：本地网络下，多媒体应用的视频质量主观评估标准
- P. 911：本地网络下，多媒体应用的视听质量主观评估标准
- P. 913：移动互联网下，在不同环境、场景下的视听质量主观评估标准
- P. 930：如何生成各种折损视频的标准

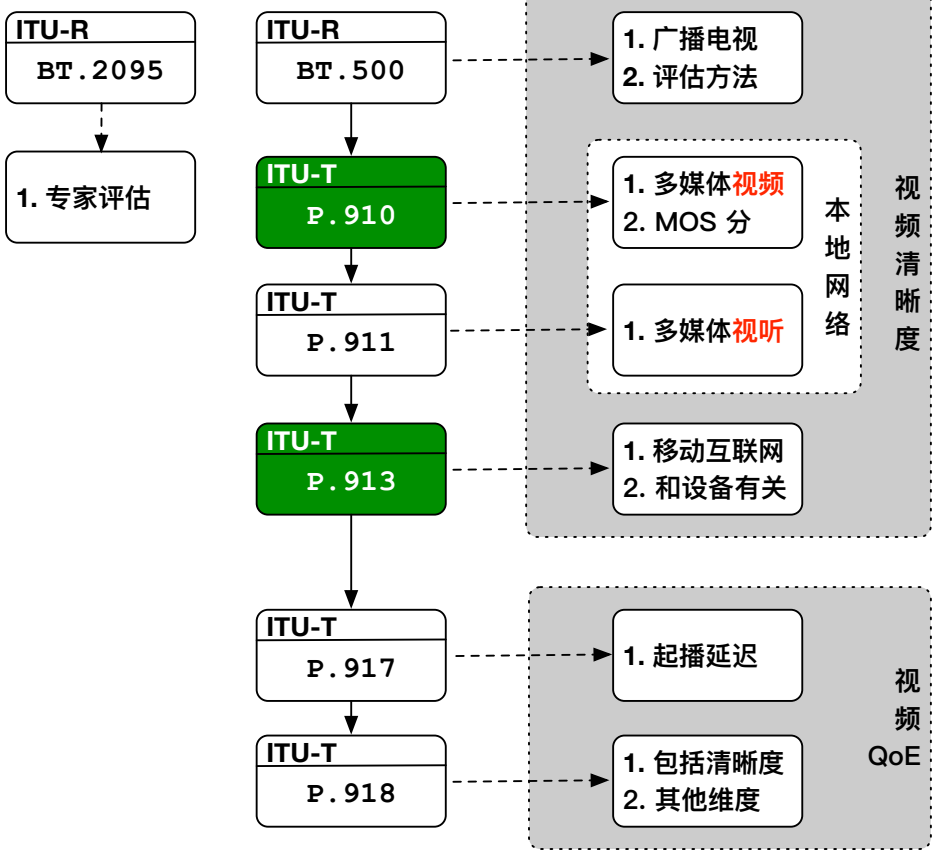
视频 QoE 主观评估标准

- P. 917：起播延迟对视频 QoE 带来的影响的主观评估方法
- P. 918：多维度下评估视频 QoE 的主观评估标准

其他类型的视频质量主观评估标准

- P. 912：用于实体识别的视频主观质量评估方法
- P. 914-916：和 3D 视频相关的评估方法
- P. 919：全景视频相关的评估方法
- P. 920/P. 931：交互式视频应用相关的评估方法，比如直播。

3. ITU 提出的视频质量主观评估体系和规范



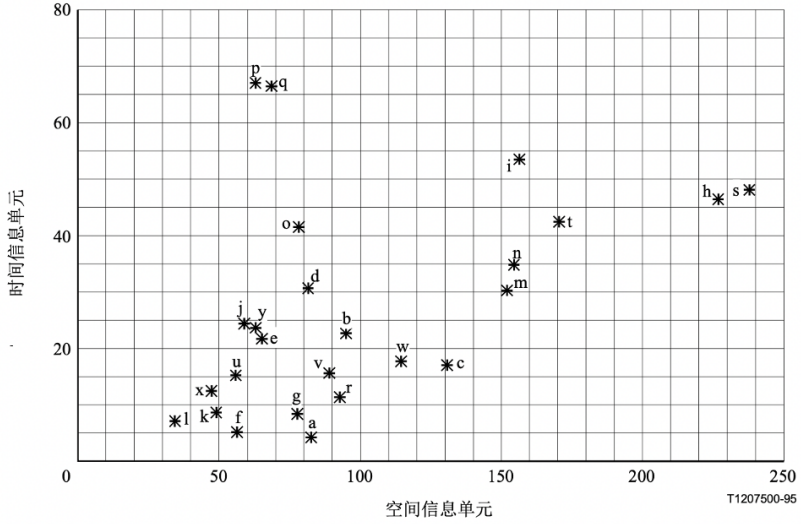
	ITU-T P.910	ITU-T P.913
样本筛选和处理	时长：8S-10S SI/TI	场景个数：4-6
评估者选择	普通用户：15人 专家用户：4人	实验室环境：24人 公共环境：35人
评估方法	ACR：单个视频 DCR：参考/折损 PC/CCR：折损/折损	
打分量级表	5分制(1~5)：优秀，良好，普通，较差，差 5分制：无感知，不厌烦，轻微厌烦，厌烦，非常厌烦 7分制(-3~3)：甚差，较差，稍差，相同，……	
评估环境	实验室环境 家庭环境	公共环境 移动设备
打分置信度	PQ方法	
评估结果	MOS/DMOS	

3.1 ITU 提出的视频质量主观评估体系——样本筛选

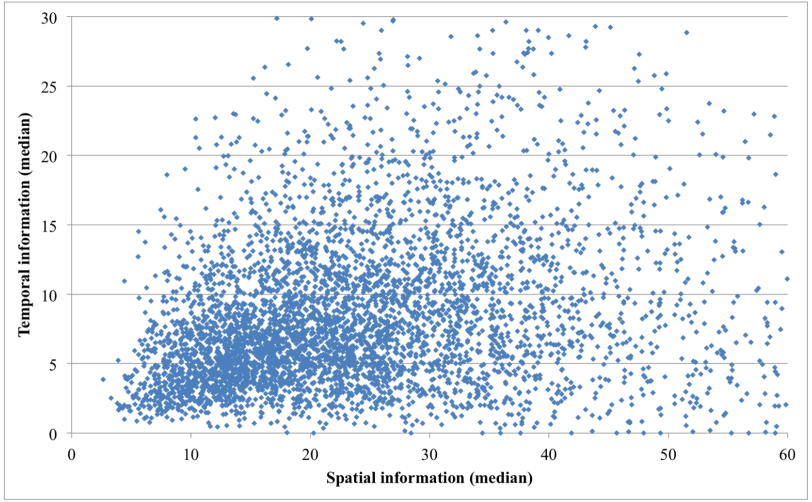
✦时长：10s

✦场景：4~6个

✦内容复杂度：SI/TI



图A.2 – 测试场景集示例的空间-时间图



(b) (SI,TI)-combinations for clips with $SI \leq 60$, $TI \leq 30$.

3.2 ITU 提出的视频质量主观评估体系——评估方法和打分指标

ACR：每次评估一个视频，然后给视频打分。

DCR：每次评估一对视频（原视频-处理后视频），给出处理后视频相对于原视频的感知打分。

PC/CCR：每次评估一对视频(系统1处理的视频-系统2处理的视频)，给出哪个视频更好的打分。

ACR: 单个视频效果分析

分数	等级	解释
5	Excellent	优秀
4	Good	良好
3	Fair	普通
2	Poor	较差
1	Bad	差

DCR: 转码效果分析

分数	等级	解释
5	Imperceptible	无感知的
4	Perceptible but not annoying	可感知但不令人厌烦
3	Slightly annoying	轻微令人厌烦
2	Annoying	令人厌烦
1	Very annoying	非常令人厌烦

PC/CCR: 竞对/编解码分析

分数	等级	解释
-3	Much worse	甚差
-2	Worse	较差
-1	Slightly worse	稍差
0	The same	相同
1	Slightly Better	稍好
2	Better	较好
3	Much Better	甚好

3.3 ITU 提出的视频质量主观评估体系——评估环境

实验室环境：有实验室环境、观看距离等的要求。**算法打分**一般都基于实验室环境产出的数据。
家庭环境：一般用的较少。
公共环境：用户使用产品的更加真实的场景，包括网络，设备，环境噪音等都有考虑。

表1 – 观测条件

参数	设置
观测距离（注1）	1-8 H（注2）
屏幕最高亮度	100-200 cd/m（注2）
非活动屏幕亮度与最高亮度之比	≤ 0.05
当在完全黑暗的屋内仅显示黑色等级时，屏幕亮度与相应的白色等级峰值之比	≤ 0.1
画面显示器背景亮度与画面亮度峰值之比（注3）	≤ 0.2
背景色度（注4）	D_{65}
屋内背景亮度（注3）	≤ 20 lux
注1 – 对于给定屏幕高度，当视觉质量劣化时，对于被试者而言较佳的观测距离可能会增长。考虑到此，进行鉴定测试前要先决定较佳的观测距离。观测距离通常取决于应用。 注2 – H表示画面高度。要根据屏幕尺寸、应用类型和实验目标来决定观测距离。 注3 – 该值表示允许最大可察觉失真的设置，对某些应用，允许具有更高值或者由应用决定。 注4 – 对PC显示器，背景色度应适应显示器色度。	

构建奈飞相关数据集

We then ran subjective tests to determine how non-expert observers would score the impairments of an encoded video with respect to the source clip. In standardized subjective testing, the methodology we used is referred to as the *Double Stimulus Impairment Scale (DSIS)* method. The reference and distorted videos were displayed sequentially on a consumer-grade TV, with controlled ambient lighting (as specified in recommendation ITU-R BT.500–13 [2]). If the distorted

抖音: 主观实景实验室

3.4 ITU 提出的视频质量主观评估体系——MOS/DMOS

MOS：平均主观得分（MOS）是主观视频质量评估中最常用的指标。

- MOS构成了主观质量评价方法的基础。
- 在电话网络领域，使用MOS评估用户对网络质量的感知已有数十年之久。
- MOS一般用于诸如ACR/ACR-HR的单刺激方法。

DMOS：差异质量评分（DMOS）和MOS类似。

- 一般用于DCR和PC/CCR方法用于比较两个视频差异的实验中。

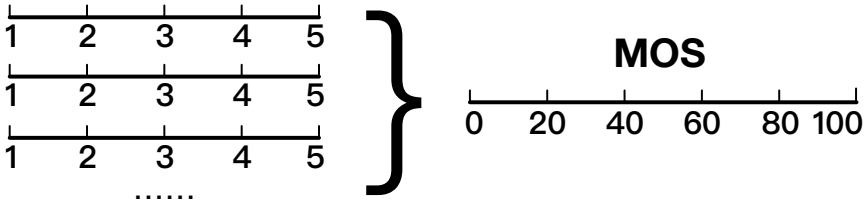
在计算MOS/DMOS时，需要确认评估者的打分是否置信，可以使用 ITU-R BT.500中附件2的第2.3.1节定义的程序筛选并去掉不置信的打分者。

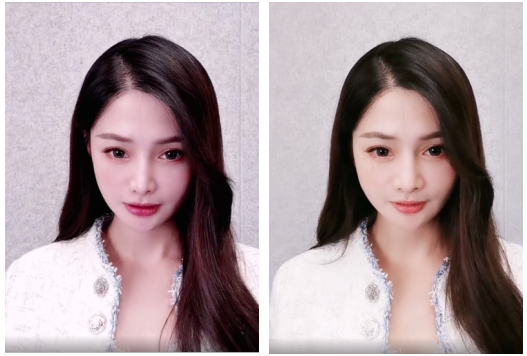
2.3.1　用于DSIS、DSCQS和替代方法的筛选，SSCQE法除外

首先用β_2测试（通过计算函数的峰态系数，即四阶动差与二阶动差平方的比值）确定测试演示的这种评分分布正常与否。如果β_2在2和4之间，则这一分布被视为正常。对于每次演示，每一观察者的评分u_{ijkr}必须与平均值$\bar{u}_{jkr}$，加上相关标准差S_{jkr}乘以2（若属正常）或乘以$\sqrt{20}$（若属异常），也就是与P_{jkr}相比较，并与相关平均值减去同样的标准差乘以2或乘以$\sqrt{20}$，也就是与Q_{jkr}相比较。每当发现观察者的评分高于P_{jkr}，与每一观察者P_i相关的计数仪就递增。同样，每当发现观察者的评分低于Q_{jkr}，与每一观察者Q_i相关的计数仪就递增。最后，必须计算下面两个比值：$P_i + Q_i$除以每一观察者在整个测试阶段内的总评分次数，以及$P_i - Q_i$除以$P_i + Q_i$得出的绝对值。如果第一个比值大于5%而第二个比值小于30%，则观察者i必须舍弃（见注1）。

4. 导致主观画质评估结论不置信的常见其他原因

1. **评估样本的问题**：时间较长、场景变化较大、原视频画质较差等都会导致评估结果的不置信。
2. **logo的问题**：评估中的样本视频含有产品logo，导致评估结果的不置信。
3. **评估设备的问题**：不同的显示设备会带来不一致的评估结论，例如：PC、手机，不同屏幕分辨率的手机之间均存在差异，大（小）屏甚至会将细节的差异放大（缩小），进而带来结果不置信。
4. **评估场景和目标场景不同的问题**：PC上评估手机上的半屏播放场景或全屏播放场景会带来结果的不置信。
5. **对MOS的认知的问题**：对于 CCR/PC 而言，50分的含义有可能不是感知不到差异，比如5个用户打2分，5个用户打4分。忽略打分的趋势而只看MOS，会带来结果的不置信。





6. **打分时关注点不一致的问题**：有的用户关注背景，有的用户关注前景，有的用户关注亮度，……
7. **单帧对比和视频对比的问题**：对于视频而言，部分帧的画质变差会因为运动的原因而忽略，这也会导致评估结果的不置信。

THANKS