国 际 电 信 联 盟

ITU-T

P.910

(04/2008)

国际电信联盟
电信标准化部门

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

多媒体业务的视听质量

多媒体应用的主观性视频质量评价方法

ITU-T P.910 建议书

ITU-T

ITU 国际电信联盟

ITU-T P 系列建议书

电话传输质量、电话装置和本地线路网络

名词术语和传输参数对用户传输质量意见的影响	系列	P.10
用户线和话机	系列	P.30
		P.300
传输标准	系列	P.40
客观测量装置	系列	P.50
		P.500
客观电声测量	系列	P.60
与话音响度有关的测量	系列	P.70
质量的客观和主观评价方法	系列	P.80
		P.800
多媒体业务的音视频质量	**系列**	**P.900**
IP端点的传输性能和业务质量问题	系列	P.1000
车辆间通信	系列	P.1100

欲了解更详细信息，请查阅ITU-T建议书目录。

ITU-T P.910 建议书

多媒体应用的主观性视频质量评价方法

摘要

ITU-T P.910建议书描述了诸如视频会议、存储和检索应用、远程医疗应用等多媒体应用单向总体视频质量评估的非交互式主观性评价方法。这些方法用于若干不同方面，包括但不限于算法的选择、音视频系统性能排名、音视频连接期间的质量水平评估。本建议书还概要说明了将使用的源序列的特性，如期限、内容类型和序列数量等。

来源

ITU-T第9研究组（2005-2008年）按照ITU-T A.8建议书规定的程序，于2008年4月6日批准了ITU-T P.910建议书。

前言

国际电信联盟（ITU）是从事电信领域工作的联合国专门机构。ITU-T（国际电信联盟电信标准化部门）是国际电信联盟的常设机构，负责研究技术、操作和资费问题，并且为在世界范围内实现电信标准化，发表有关上述研究项目的建议书。

每四年一届的世界电信标准化全会（WTSA）确定ITU-T各研究组的研究课题，再由各研究组制定有关这些课题的建议书。

WTSA第1号决议规定了批准ITU-T建议书须遵循的程序。

属ITU-T研究范围的某些信息技术领域的必要标准，是与国际标准化组织（ISO）和国际电工技术委员会（IEC）合作制定的。

注

本建议书为简明扼要起见而使用的“主管部门”一词，既指电信主管部门，又指经认可的运营机构。

遵守本建议书的规定是以自愿为基础的，但建议书可能包含某些强制性条款（以确保例如互操作性或适用性等），只有满足所有强制性条款的规定，才能达到遵守建议书的目的。“应该”或“必须”等其它一些强制性用语及其否定形式被用于表达特定要求。使用此类用语不表示要求任何一方遵守本建议书。

知识产权

国际电联提请注意：本建议书的应用或实施可能涉及使用已申报的知识产权。国际电联对无论是其成员还是建议书制定程序之外的其它机构提出的有关已申报的知识产权的证据、有效性或适用性不表示意见。

至本建议书批准之日止，国际电联尚未收到实施本建议书可能需要的受专利保护的知识产权的通知。但需要提醒实施者注意的是，这可能并非最新信息，因此特大力提倡他们通过下列网址查询电信标准化局（TSB）的专利数据库：http://www.itu.int/ITU-T/ipr/。

目录

页码

ITU-T P.910 建议书

多媒体应用的主观性视频质量评价方法

1 范围

本建议书旨在定义非交互式主观性评价方法，用于评估以TV3、MM4、MM5和MM6类别中指定的比特率编码的数字视频图像的质量，如表D.2中所述，用于诸如视频电话、视频会议、存储及检索等应用。可将这些方法用于若干不同方面，包括但不限于算法的选择、视频系统性能排名、视频连接期间的质量水平评估。

2 参考文献

下列ITU-T建议书和其他参考文献的条款，在本建议书中的引用而构成本建议书的条款。在出版时，所指出的版本是有效的。所有的建议书和其它参考文献均会得到修订，本建议书的使用者应查证是否有可能使用下列建议书或其它参考文献的最新版本。当前有效的ITU-T建议书清单定期出版。本建议书引用的文件自成一体时不具备建议书的地位。

[ITU-T J.61]	ITU-T J.61建议书（1988年），为国际连接而设计的电视电路的传输性能。
[ITU-T P.800]	ITU-T P.800建议书（1996年），传输质量的主观性测定方法。
[ITU-T P.930]	ITU-T P.930建议书（1996年），视频参考损害系统的原则。
[ITU-R BT.500-9]	ITU-R BT.500-9建议书（1988年），电视画面质量的主观性评价方法。
[ITU-R BT.601-4]	ITU-R BT.601-4（1994年），演播室的数字电视编码参数。
[ITU-R BT.814-1]	ITU-R BT.814-1（1994年），用于设置显示器亮度和对比度的规范及校准程序。
[IEC/TR 60268-13]	IEC/TR 60268-13 (1998), *Sound system equipment – Part 13: Listening tests on loudspeakers* <http://webstore.iec.ch/webstore/webstore.nsf/artnum/022890>.

3 术语和定义

本建议书规定下列术语：

3.1 γ值（gamma）：描述视觉显示的灰度级之间区别的参数。屏幕亮度和输入信号电压之间的关系是非线性的，电压按一个指数γ提升。为了补偿这个非线性，通常在摄像机采用一个γ的反函数。γ也会影响颜色的重现。

3.2 优化测试（optimization tests）：典型地，在新算法或系统进行开发或标准化时完成的主观性测试。这些测试的目标是评估新工具的性能以便优化正在研究的算法或系统。

3.3 鉴定测试（qualification tests）：为了比较大批量生产的系统或设备而实施的典型主观性测试。这些测试必须在尽可能代表实际使用条件的测试条件下完成。

3.4 空间感觉信息（spatial perceptual information (SI)）：通常指示图片的空间细节总量的度量参数。通常空间场景越复杂它就越高。它不意味着对平均信息量的度量也与通信理论定义的信息无关。SI的方程式参见5.3.1小节。

3.5 时间感觉信息（temporal perceptual information (TI)）：通常指示视频序列上时间变化总量的度量。通常序列运动愈高它就愈高。它不意味着对平均信息量的度量也与通信理论定义的信息无关。TI的方程式参见5.3.2小节。

3.6 透明（保真）（transparency (fidelity)）：一个说明编解码器或系统的性能相当于没有任何劣化的理想传输系统的概念。

能够定义两类透明：

第一类采用数学准则说明被处理的信号是多么完美地符合输入信号，或理想信号。如果没有差异，系统就是完全透明。第二类由试验观测者说明被处理的信号与输入信号或理想信号吻合程度如何。如果在任何试验条件下都感觉不到差异，该系统感觉上是透明的。没有明确指明参考准则的术语“透明”被用于感觉上透明的系统。

3.7 复制（replication）：再现相同主题的相同电路状态（用相同来源的材料）。

3.8 主观性测试的信度（reliability of a subjective test）：

a) 个体内（“被试者本身”）的信度是指某一个被试者对同一测试状态重复得出的评分的一致性。

b) 个体之间（“被试者之间”）的信度是指对同一测试状态各个被试者得出的评分的一致性。

3.9 主观性测试的有效性（validity of a subjective test）：测试得到的平均评分与由测量表明的真实值之间的一致性。

3.10 参考条件（reference conditions）：为使不同的试验得出的评估稳定，附加在测试条件上的虚拟条件。

3.11 显参考（源参考）（explicit reference (source reference)）：在采用DCR方法时，评价人使用的，作为表达他们意见的参考状态。在每一序列对内首先呈现此参考。通常，显参考的格式是用在被测编解码器输入的格式（例如，[ITU-R BT.601-4]建议书种定义的CIF、QCIF、SIF等）。在本建议书的正文中，当上下文已经明确“参考”的含义时，省略“显”和“源”这两个字。

3.12 隐参考（implicit reference）：在采用ACR方法时，评价人使用的、作为表达他们对测试材料的意见的参考状态。如果隐参考是由试验者提出的，它必须是全部评价人都熟知的（即，常规电视系统、实体）。

4 缩略语

本建议书使用下列缩略语：

ACR 绝对等级评分

ACR-HR 带有隐参考的绝对等级评分

CCD 电荷耦合器件

CI 置信区间

CIF 通用媒体模式

注 – [b-ITU-T H.261]为视频电话定义的图片格式为：352行×288像素。

CRT 阴极射线管

DCR 劣化等级评分

DV 差异观测者

%GOB 好或更好的百分比（良好和优秀的比例）

LCD 液晶显示屏

MOS 平均意见评分

PC 配对比较

%POW 差或更差的百分比（较差和差的评分比例）

PVS 处理后的视频序列

QCIF 四分之一CIF

注 – [b-ITU-T H.261]为视频电话定义的图片格式为：176 行×144像素。

S/N 信号噪声比

SI 空间信息

SIF 标准中间格式

注 – [b-ISO/IEC 11172]（MPEG-1）中定义的图片格式为：352 行×288像素×25帧/s和352行×240像素×30 帧/s。

SP 同步呈现

std 标准偏差

TI 时间信息

VTR 磁带录像机

5 源信号

为了控制源信号的特性，应根据测试和目标来定义测试序列，并记录在数字存储系统。当实验者有兴趣对来自不同实验室的结果进行比较时，需要使用一组公共的源序列，以消除进一步的变化源。

5.1 录制环境

可将光源（灯泡或荧光灯）放置在照相机上方或侧面。放置光源时，办公室照明是较典型的开销，应与描绘商业环境的场景一起使用。同时，应避免演播室照明和其他非典型光源。

室内使用的房间照明条件变化范围在100 lux至大约10 000 lux。必须考虑光（荧光灯照明）的变化（交流频率），因为这可能造成录制的视频序列中出现闪烁。

应该仔细控制并报告照明条件、墙体颜色、表面反射率等。

5.2 录制系统

5.2.1 照相机

应使用高品质的CCD照相机记录图片序列。

输入视频信号的信噪比对编解码器性能的影响很大。

为了定义视频输入，应明确下述关键点：

- YUV信号的动态范围；
- γ校正因子（应为0.45）；
- 带宽/滤波器的斜率；
- 在很差的照明条件下照相机的灵敏度以及自动增益控制（AGC）的特性（如果使用）。

加权S/N应根据[ITU-T J.61]建议书3.2.1小节C部分的定义来测量。加权S/N应大于45 dB r.m.s。

时钟信号的不稳定或抖动可能会带来噪音影响。照相机计时装置需要最低0.5 ppm的稳定度。

可以使用固定或者可变焦距系统。对于台式机终端，合理的焦距深度在30 cm至120 cm的范围内，而对于多用户系统，合理的焦距深度应大于50 cm。为了支持录制室的照明度变化，应使用可调节的虹膜或者中性密度滤波器。照相机有自动白平衡功能，因此可以实现对光源色温的适应。白色色温的校正范围可以从2700° K（室内使用电灯泡）至6500° K（多云的白天温度）。

5.2.2 视频信号和存储格式

照相机提供的视频源信号应按照[ITU-R BT.601-4]建议书中的A部分进行采样。为了避免源信号失真，应以数字格式存储，例如，用计算机或D1 4:2:2磁带格式存储。

5.3 场景特征

选择测试场景非常重要。场景的空间和时间感知信息尤其是重要参数。这些参数在决定可能的视频压缩总量时发挥关键性作用，因此，通过固定速率数字传输业务信道发送时，场景会受到损害。必须选择普通和相关的视频测试场景，使得其空间和时间信息与数字传输业务信道旨在提供的视频业务的信息一致。测试场景应该覆盖被测设备的用户所感兴趣的空间和时间信息的全部范围。

附件A和附录I、II给出了有关测试序列特性的细节以及合适的测试场景示例。

序列的数量应根据实验设计来决定。为了避免观测者感到无聊并实现测试的最低信度，至少应为序列选择四种不同类型的场景（即：不同主题的场景）。

下文的几个小节展示了量化测试场景的空间和时间信息的方法。这些用于评估测试场景空间和时间信息的方法在现在和未来都适用于视频质量测试。在空间-时间矩阵内的视频场景的位置很重要，因为发送的视频场景（尤其通过低编码速率的编解码器）的质量很大程度上取决于该位置。这里呈现的空间和时间信息测量可以用于保障空间-时间平面的合理覆盖。

以下给出的空间和时间信息测量是为完整测试序列上的各个帧单独赋值。在时间序列值中，该结果通常将在某种程度上有所变化。以下给出的感知信息测量用最大函数（序列的最大值）消除了这种可变性。可以对可变性本身开展有益的研究，例如，逐帧形式的空间-时间信息图。在测试序列上使用信息分发也会使得可使用场景剪辑对场景进行更好的评估。

5.3.1 空间感知信息测量

空间感知信息（SI）基于Sobel滤波器。在时间n（F_n），各个视频帧（亮度平面）首先用Sobel滤波器[$Sobel(F_n)$]进行滤波。然后，计算各个经Sobel滤波器滤波后的帧的像素的标准差（std_{space}）。为视频序列中各个帧重复该操作，产生场景的空间信息的时间序列。选择时间序列（$\max_{time}$）中的最大值，以表示场景的空间信息内容。该过程可以以如下所示的方程式的形式表示：

$$SI = \max_{time}\left\{std_{space}\left[Sobel(F_n)\right]\right\}$$

5.3.2 时间感知信息测量

时间感知信息（TI）基于运动差异特性，$M_n(i, j)$，其是处于空间内相同位置的连续时间或帧的像素值（亮度平面）差异。作为时间(n)的函数$M_n(i, j)$的定义如下：

$$M_n(i,j) = F_n(i,j) - F_{n-1}(i,j)$$

其中，F_n（i， j）是时间上第n帧第i行和第j列处的像素。

通过计算所有i和j的$M_n(i, j)$空间（std_{space}）上的标准差的最大时间值（$\max_{time}$），来实现时间信息（TI）的测量。

$$TI = \max_{time}\left\{std_{space}\left[M_n(i,j)\right]\right\}$$

相邻帧的更多运动会导致更高的TI值。

注 – 对于包括场景剪辑的场景，可能会给出两个值：一个是场景剪辑包含在时间信息测量内，一个是场景剪辑不在时间信息测量内。

6 测试方法和实验设计

图像的感知质量测量需要使用主观性缩放方法。这些测量成立的条件是“刺激”的物理特性间存在相关性，在这种情况下，在测试中呈现给被试者的视频序列，以及由刺激引起的感觉的量级和性质可被测量。

许多实验方法已经被验证可用于不同的目的。对于使用TV3、MM4、MM5和MM6类别中指定的比特率的应用，建议使用此处建议的三种方法，如表D.2所示。附录III和IV中描述了更多测试方法。

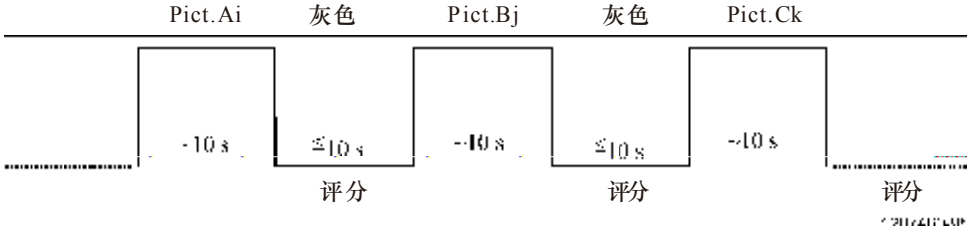
为特定应用选择一种测试方法取决于许多因素，诸如：环境、目的、以及在发展过程的哪个阶段开展测试等。

6.1 绝对等级评分（ACR）

绝对等级评分方法是一种对测试序列一次呈现一个，且在等级量表内进行独立评分的等级判断。（这种方法也称为单个刺激法。）

该方法明确了在每次呈现之后，被试者都要评估展示的序列的质量。

图1说明了刺激呈现的时间模式。如果采用连续的评分时间（例如，几个观测者同时从磁带上进行评分），评分时间应该小于等于10 s。根据不同的测试材料内容，呈现时间可能会有所增减。



Ai 测试条件i下的序列A
Bj 测试条件j下的序列B
Ck 测试条件k下的序列C

图1 – ACR方法中的刺激呈现

使用下列5种量级评价总体质量：

5 优秀

4 良好

3 普通

2 较差

1 差

如果需要更高的识别能力，可以使用九级量表。附件B给出了合适数值或连续量级的示例。附件B还给出了总体质量以外的评分维度示例。当测试环境下一些系统的总体质量评分基本相同（即使这些系统可被清楚地感知为不同）时，这些维度可用来获取不同感知质量因素的更多信息。

对于ACR方法，通过在测试中的不同时间点重复相同的测试条件来获得所需的复制次数。

6.2 带有隐参考的绝对等级评分（ACR-HR）

带有隐参考的绝对等级评分方法绝对是一种对测试序列一次呈现一个，且在等级量表内进行独立评分的等级判断。目前的测试步骤必须包括出现作为任意其他测试刺激的每个测试

序列的参考版本。这被称为隐参考条件。在进行数据分析时，将计算每个测试序列和其对应的（隐）参考间的差异质量评分（DMOS）。这个步骤就是“隐参考”。

该方法明确了在每次呈现后，被试者都被要求评估展示的序列的质量。

图1展示了刺激呈现的时间模式。如果采用连续的评分时间（例如，几个观测者同时从磁带上进行评分），评分时间应该小于等于10 s。根据不同的测试材料内容，呈现时间可能会有所增减。

使用下列5种量级评价总体质量：

5 优秀

4 良好

3 普通

2 较差

1 差

差异观测者分数（DV）在每个被试者/每个处理后的视频序列（PVS）的基础上计算得出。通过下述公式，合适的隐参考（REF）被用于计算DV：

$$DV(PVS) = V(PVS) – V(REF) + 5$$

其中V是观测者的ACR分数。使用此公式时，DV为5时表示质量“优秀”，为1时表示质量“差”。任何大于5的DV值（即，处理后的序列评分比其相关的隐参考序列高）通常情况下被认为是有效的。否则，可应用两点粉碎功能来预防这些单独的ACR-HR观测者分数（DV）过度影响总体平均意见分数：

$$当DV > 5时，crushed\_DV = (7*DV)/(2+DV)$$

如果需要更高的识别能力，可以使用九级量表。附件B给出了合适数值或连续量级的示例。附件B还给出了总体质量以外的评分维度示例。当测试环境下一些系统的总体质量评分基本相同（即使这些系统可被清楚地感知为不同）时，这些维度可用来获取不同感知质量因素的更多信息。

对于ACR方法，通过在测试中的不同时间点重复相同的测试条件来获得所需的复制次数。

ACR-HR方法只能与业界专家认为在上述五级量表中质量“良好”或“优秀”的参考视频一起使用。

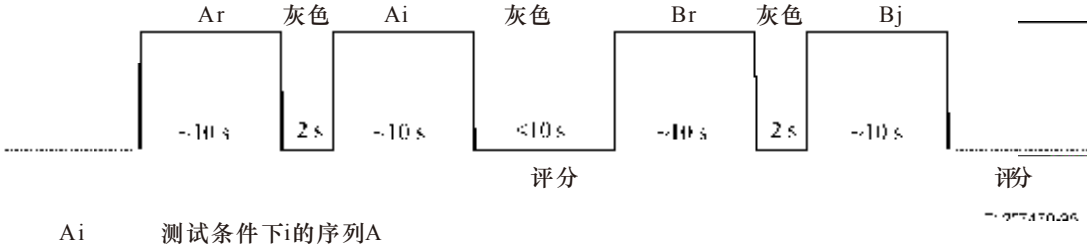
ACR-HR方法不适用于分析发生在视频序列最初或者最后一秒的异常损害。观测者不熟悉参考视频序列可能会造成其他明显的损害被忽略（例如，如果序列在结束之前立即暂停，观测者可能无法确定这是故意的行为还是网络错误）。

6.3 劣化等级评分（DCR）

劣化等级评分揭示了测试序列成对呈现：每对中第一个出现的刺激通常是源参考，第二个刺激是通过被测系统呈现的相同源。（这种方法也被称为双刺激损害量表法。）

当时用压缩的图片格式时（例如：CIF、QCIF、SIF），在同一个显示器上同时显示参考和测试序列是有用的。附件C讨论了关于此呈现步骤的说明。

图2展示了刺激呈现的时间模式。如果采用连续的评分时间（例如，几个观测者同时从磁带上进行评分），评分时间应该小于等于10 s。根据不同的测试材料内容，呈现时间可能会有所增减。



Ai　　测试条件下i的序列A
Ar,Br　　使用参考源格式的序列A和B
Bj　　测试条件下j的序列B

图2 – DCR方法中的刺激呈现

在这种情况下，要求被试者对与参考相关的第二次刺激的损害进行评分。

使用下列5种量级评价损害：

5　　无感知的

4　　可感知但不令人厌烦

3　　轻微令人厌烦

2　　令人厌烦

1　　非常令人厌烦

对于DCR方法，通过在测试中的不同时间点重复相同的测试条件来获得所需的复制次数。

6.4　　配对比较法（PC）

配对比较法揭示了测试序列成对呈现，包括首先通过一个被测系统然后通过另一个系统呈现的相同序列。

被测系统（A、B、C等）通常用n种（n – 1）可能的组合形式进行组合，如AB、BA、CA等。因此，所有序列对都能以可能的顺序（例如：AB、BA）显示。这就要求观测者在每对序列出现后对在测试场景下这一对序列中哪个要素是首选做出判断。

图3展示了刺激呈现的时间模式。如果采用连续的评分时间（例如，几个观测者同时从磁带上进行评分），评分时间应该小于等于10 s。呈现时间应约为10 s，且根据不同的测试材料内容，呈现时间可能会有所增减。

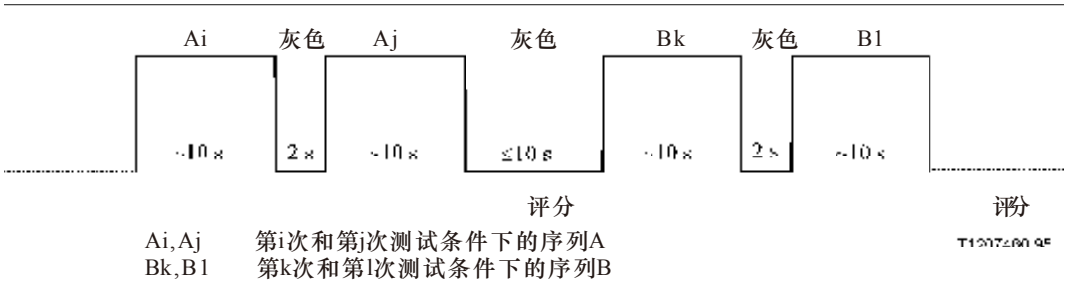


图3 – PC方法中的刺激呈现

当采用低分辨率时（例如：CIF、QCIF、SIF），应在同一个显示器上同时显示每个序列对。附件C讨论了关于此呈现步骤的说明。

对于PC方法，通常不需要考虑复制次数，因为该方法自身揭示了相同条件下不同序列对的重复呈现。

一种PC方法的变体利用等级量表来进一步测量序列对之间的差异。参见[ITU-R BT.500-9]建议书和[ITU-T P.800]建议书。

6.5 方法比较

选择测试方法时，有一个重要的问题是使用显参考（例如：DCR）的方法和不使用显参考的方法（例如：ACR、ACR-HR和PC）之间的根本差异。第二类方法不测试透明或保真。

当测试源信号传输的保真时，应使用DCR方法。在评估高质量系统时，这通常是一个重要因素。在[ITU-R BT.500-9]建议书中明确的DCR长久以来作为评价电视画面的主要方法，其典型质量代表了极高等级的视频电话和视频会议。也可用其他方法来评估高质量系统。当观测者对于损害的检测是一项重要因素的时候，对于DCR量表（不可感知的/可感知的）的具体评价是很具价值的。

因此，当测试源信号传输的保真时，应使用DCR方法。

在多媒体通信环境下，也可用DCR方法来评估高质量系统。DCR量表中支持不可感知/可感知的损害的区分，以及参考质量的比较。

ACR可方便快速地实施，且刺激的呈现与系统的正常操作相似。因此，ACR很适合鉴定测试。

ACR-HR方法是拥有ACR方法在呈现和速度方面的所有优势。而其相比于ACR，还有明显的优点，其中主要的是可从主观性评分中移除参考视频的感知影响。这减少了场景偏差的影响（例如，观测者喜欢或者不喜欢参考视频）、参考视频质量（例如，照相机质量的微小差异）和显示最终评分结果的显示器（例如，专业质量与消费者等级）。ACR-HR很适合大型实验，但要在所有的参考视频都至少达到质量“良好”的前提下。然而，ACR-HR对于一些用直接差异方法（例如，DCR）可轻易检测的损害并不敏感。例如，ACR-HR可能无法检测颜色增益（例如，暗淡的颜色）的系统性衰减。

PC方法的主要优点是其高识别能力，当几个测试项目的质量几乎相等时，这是特别有价值的。

当同一测试中有大量项目需要评估时，基于PC方法的步骤可能会较长。在这种情况下，首先可以用有限数量的观测者进行ACR或DCR测试，其次对于那些收到相同评分的项目单独进行PC测试。

6.6 参考条件

质量评价的结果通常不仅取决于实际视频质量，也取决于诸如测试条件的总体质量范围、评估者的经验和预期等其他因素。为了控制这样的影响，可以添加一些虚拟测试条件并将其用作参考。

[ITU-T P.930]建议书给出了参考条件和产生这些条件的步骤。当测试项目引入的损害较小时，特别推荐在PC测试中引入源信号作为参考条件。

参考条件的质量等级应至少覆盖测试项目的质量范围。

6.7 实验设计

可将[b-Kirk]用于不同的实验设计，诸如完全随机设计、拉丁、希腊-拉丁方和尤登方设计、复制块设计等，应根据试验的目的来选择合适的设计。

接下来，实验者要选择实验的设计方法，以实现特定的成本和精度目标。设计还可以取决于给定测试中的特定兴趣点。

建议在实验中至少包括两种、如果可能三种或四种复制（即，再现相同的条件）。使用复制有很多原因，最重要的是，“被试者内部的变化”可使用复制的数据来测量。为了测试被试者的信度，可在相同条件下使用相同顺序的呈现。如果使用了不同顺序的呈现，由此产生的实验数据的变化由顺序效应和被试者内部的变量组成。

复制使计算被试者的信度变得可能，如果必要，可放弃来自某些被试者的不可靠的结果。在被试者标准偏差之内和之间的估计值是进一步进行正确的差异分析并将结果推广到更广泛人群的先决条件。此外，复制确保测试中的学习效果在某种程度上能够得以平衡。

学习效果的进一步改善通过包括一个训练阶段来获得，其在每个测试阶段开始时至少呈现5个条件。阶段后期，这些条件可被选定作为具有代表性的呈现而展示。在对测试结果进行统计分析过程中，不考虑初步呈现。

7 评估程序

表1列出了视频质量评价的典型观测条件。应明确评价中使用的实际参数设置。为了进行测试结果的比较，所有的观测条件必须固定且同类测试的实验室都具备相同条件。

显示器的尺寸和类型应符合被研究的应用。当序列通过基于PC的系统呈现时，必须明确显示器的特征，例如：显示器的点距、使用的视频显示卡的类型等。

关于显示格式，最好使用全屏显示序列。然而，由于某些原因，当序列必须显示在屏幕的窗口上时，屏幕的背景颜色应为对应于Y=U=V=128（U和V无符号）的50％灰色。

7.1 观测条件

应该在下述参考条件下进行测试：

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。	

7.2 处理和重放系统

从源记录中获取测试图像有两种方法：

a) 通过被测系统实时传输或重播视频录像，被试者同时观看和响应；

b) 通过离线处理被测设备的源记录，记录输出，并给出一组新的记录。

在第二种情况下，数字磁带录像机会最小化录制过程中所产生的损害。在任何一种情况下，都要考虑由低比特率编码方案所带来的损害通常比由调制带来的损害更加明显，因此应使用诸如D2、MII和BetacamSP等专业质量数字磁带录像机。

在使用阴极射线管、液晶显示屏、等离子、投影或其他任何类型的显示器时，都要考虑应用的类型和实验目标。使用的显示器的尺寸和类型都要符合被研究的应用。

显示器应根据[ITU-R BT.814-1]建议书中规定的程序进行校准。

7.3 观测者

观测测试中（以及终端或服务的可用性测试）可能的被试者数量为4至40个。出于统计的原因，4是绝对最小值，而超过40也几乎没有意义。

特定测试中被试者的实际数量取决于所需的有效性以及将样本推广到更大的人群的需要。

一般情况下，至少应有15名观测者参与实验。直接参与图片质量评估不是其工作的一部分，他们也不应该是经验丰富的评估者。

然而，在视频通信系统发展的早期阶段以及在较大型的实验之前所进行的试点试验中，专家小组（4至8人）或其他关键性的被试者可给出指示性结果。

项目开始前，观测者通常应进行正常视力筛选或被矫正到正常的视力和视觉。关于视力，应保证观测者在标准视力表[b-Snellen]的20/30行观测无误。视力表应按照测试观测距离进行缩放，且视力测试应在观测视频图像的同一位置进行（即，将视力表靠在显示器上），然后让被试者坐下。关于视觉，观测者观测12个图版[b-Beck]错误不应多于2个。

7.4 对观测者和训练项目的说明

实验开始前，应为被试者提供被测系统的预期应用场景。此外，关于评价类型的描述、意见量表以及刺激的呈现应以书面形式给出。应在初步试验中呈现损害的范围和类型，其可能包括在实际测试中除视频序列以外的其他序列。

不能暗示在训练集中看到的最低质量必须对应于量表中的最低等级。

应谨慎回答关于步骤的问题和说明的意义，避免误解的产生，同时只能在项目开始前给出这样的解释。

提供给评估者的可能性说明文本在附录II中给出。

8 统计分析和结果报告

应将结果与实验计划的细节一起报告。对于测试变量的每一种组合，应给出评价等级的平均值和统计分布的标准偏差。

数据应包括被试者的信度计算并报告评价被试者信度所用的方法。关于被试者信度的一些标准在[ITU-R BT.500-9]建议书和[IEC/TR 60268-13]中给出。

分析累积评分分布是有益的。因为累积分布不是线性敏感的，这对于线性可疑的数据（如通过使用ACR和DCR方法获得的数据以及没有分级的等级量表（即，等级判断））而言可能特别有用。

用于ACR的数据可以用表2所示的示例来组织。

表2 – 用于ACR方法的具有累积分布评分的信息表

条件	总评分	优秀	良好	普通	较差	差	MOS	CI	Std	%GOB	%POW
条件：标记表示测试变量的组合。 **总评分**：在此条件下搜集的评分数。 **优秀、普通…差**：每次评分的出现。											

应用经典的方差分析技术来评估测试参数的意义。如果评价旨在根据参数来评估视频质量，则曲线拟合技术对于数据的解释可能是有用的。

在配对比较的情况下，计算区间尺度中每个刺激的位置的方法（其中，刺激之间的差异对应于优先权的差异），在《通话计时手册》，[b-ITU-T手册]2.6.2C小节中描述。

附件A

与测试序列特性描述相关的细节

（此附件是本建议书不可分割的组成部分）

A.1 Sobel滤波器

Sobel滤波器是通过在视频帧上卷积两个3×3内核并采用这些卷积结果的平方和的平方根来实现的。

对于y = Sobel(x)，令$x(i, j)$表示第i行和第j列处输入图像的像素。$Gv(i, j)$将是第一个卷积的结果，并以下述公式给出:

$$\begin{aligned} Gv(i,j) = & -1\times x(i-1,j-1)-2\times x(i-1,j)-1\times x(i-1,j+1)+ \\ & +0\times x(i,j-1)+0\times x(i,j)+0\times x(i,j+1)+ \\ & +1\times x(i+1,j-1)+2\times x(i+1,j)+1\times x(i+1,j+1) \end{aligned}$$

相似地，$Gh(i, j)$为第二个卷积的结果，并以下述公式给出:

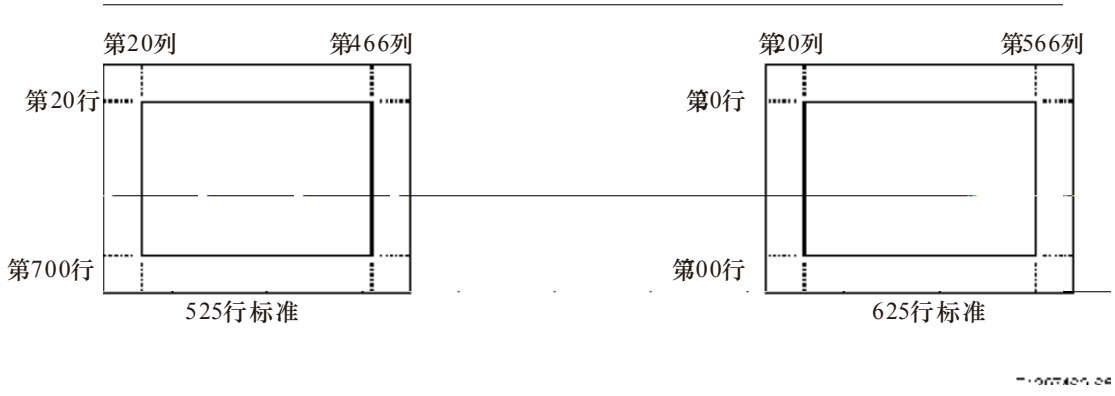
$$\begin{aligned} Gh(i,j) = & -1\times x(i-1,j-1)+0\times x(i-1,j)+1\times x(i-1,j+1)+ \\ & -2\times x(i,j-1)+0\times x(i,j)+2\times x(i,j+1)+ \\ & -1\times x(i+1,j-1)+0\times x(i+1,j)+1\times x(i+1,j+1) \end{aligned}$$

因此，第i行和第j列的Sobel滤波图像的输出为：

$$y(i,j)=\sqrt{[Gv(i,j)]^2+[Gh(i,j)]^2}$$

对所有$2 \le i \le N-1$和$2 \le j \le M-1$进行计算，其中N是行数，而M是列数。

建议在视频帧的子图像上进行该计算，以避免多余的边际效应，同时也由于视频帧的极边对于CRT用户是不可见的。这可以通过使用图A.1中所示的合适的子图像来实现，例如625和525行[ITU-R BT.601-4]格式。



图A.1 – [ITU-R BT.601-4]525行和625行格式下用于计算SI和TI的子图像

可从[b-Gonzalez]中找到更多关于Sobel滤波器的信息。

A.2 如何将SI和TI用于测试序列选择

选择测试序列时，比较不同可用序列中出现的相关空间信息和时间信息是有用的。一般而言，压缩难度与序列的空间和时间信息直接相关。

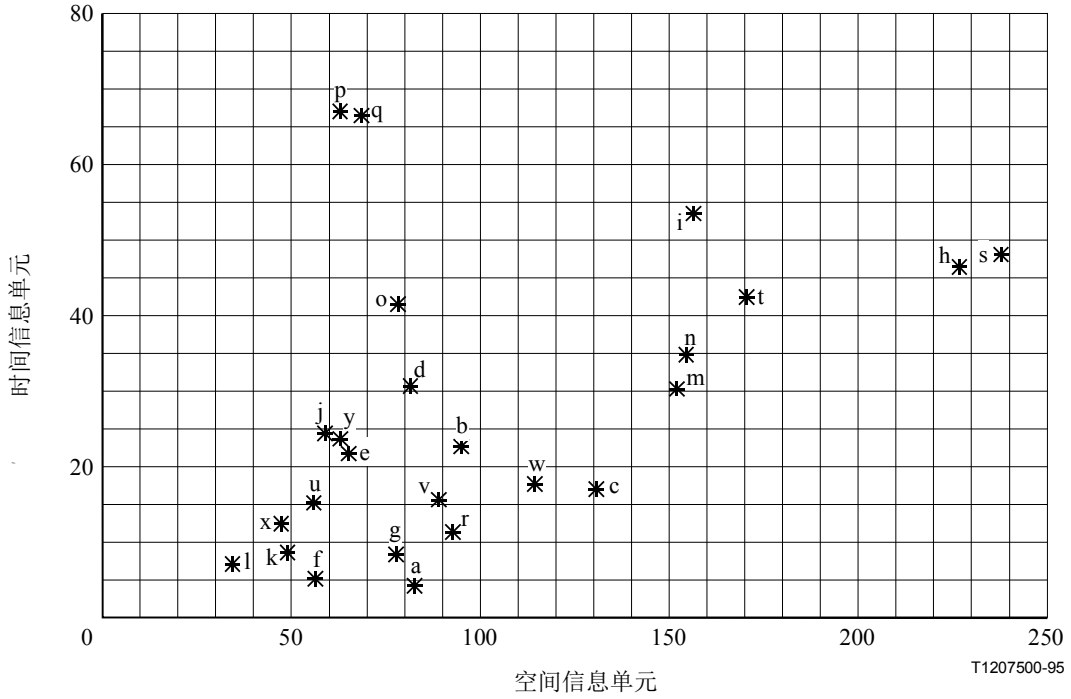
如果在给定测试中将使用数量较少的测试序列，选择跨越大部分空间-时间信息平面（如图A.2所示）的序列很重要。如果在一个测试中将使用4个测试序列，可分别从空间-时间信息平面的4个象限中选择每个序列。

相反地，如果尝试选择编码难度相当的测试序列，则选择具有相似的SI和TI值的序列。

A.3 示例

图A.2显示了一些代表性测试场景的空间和时间信息的相对量以及它们如何被放置在空间-时间信息平面上。

沿TI = 0轴（沿图的底部）可发现静止场景和具有有限运动（如l、f和a）的场景。接近图的顶部可发现具有大量运动（诸如p、q和i）的场景。沿着SI = 0轴（在图的左边）可发现具有最小空间细节（诸如l、k、x、u和f）的场景。靠近图右边可发现具有最大空间细节（诸如h和s）的场景。通过使用上述方程式及根据[ITU-R BT.601-4]规范进行空间采样的视频来获取SI和TI值。表A.1列出了按照场景内容类别分类的测试场景示例。



图A.2 – 测试场景集示例的空间-时间图

表A.1 – 场景内容类别

类别	描述	场景名称和字母
A	一个人，主要是头部和肩膀，有限的细节和运动	vtc1nw(f), susie(j), disguy(k), disgal(1)
B	一个有图形和/或更多细节的人	vtc2mp(a), vtc2zm(b), boblec(e), smity1(m), smity2(n), vowels(w), inspec(x)
C	超过一个人	3inrow(d), 5row1(g), intros(o), 3twos(p), 2wbord(q), split6(r)
D	图形指向	washdc(c), cirkit(s), rodmap(t), filter(u), ysmite(v),
E	高的对象和/或照相机运动（广播电视的示例）	flogar(h), ftball(i), fedas(y)

附件B

额外的评估量表

（此附件是本建议书不可分割的组成部分）

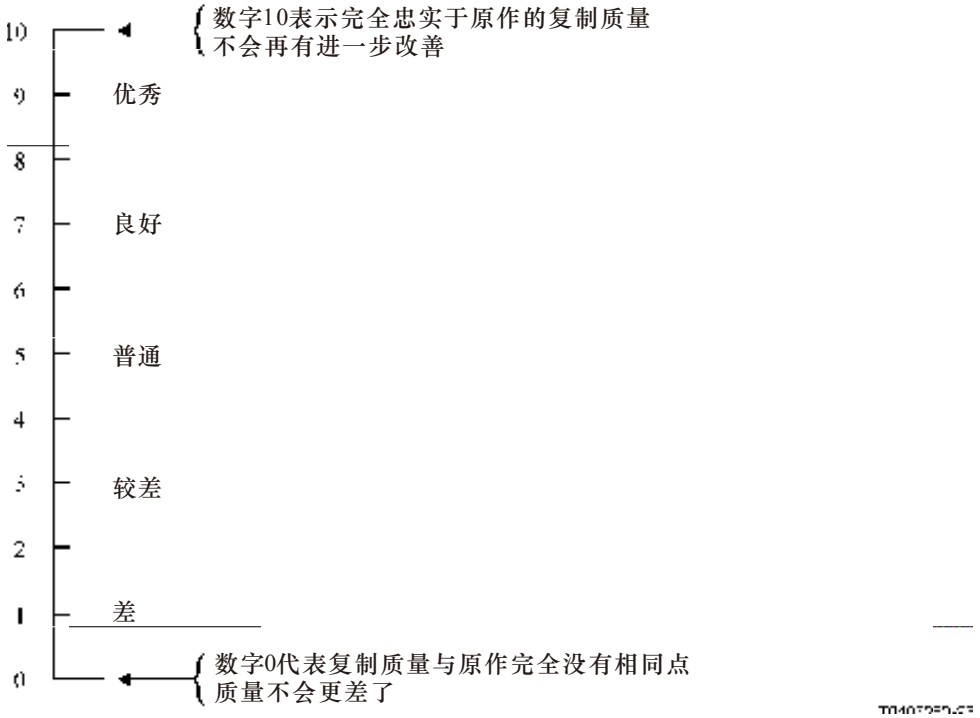
B.1 等级量表

专用于低比特率视频编解码器的评价，通常需要使用多于五级的等级量表。出于此目的，合适的量表是九级量表，如第6.1节所述，将5个口头定义的质量等级作为量表内每个第二等级的标记，如图B.1所示。

9 优秀
8
7 良好
6
5 普通
4
3 较差
2
1 差

图B.1 – 九级数值质量量表

图B.2展示了本量表的进一步延伸，端点已被口头定义为不用于评级的锚定点。在此口头定义中，使用了一些参考（例如，在图B.2中，将原始的用作参考）。这些参考可以是显或隐的，其将在训练阶段被明确阐述。也可参见[IEC/TR 60268-13]和[b-ITU-手册]2.6节（量级a）。



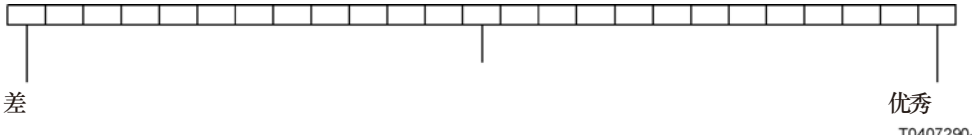
图B.2 – 十一级数值质量量表

对于两种类型的量表，被试者的响应都应该以数值（在响应表格上以书面的形式记录）或在量表上标记（在这种情况下，对于每一种评分条件，都应给被试者一个单独的量表）的形式记录。当需要数值响应时，被试者应被鼓励使用小数形式（例如，2.2而不是2），但其也可以选择只使用整数形式。

值得注意的是，将量表等级的名字翻译成各种语言比较困难。在这种情况下，等级之间的关系可能会变得与原始语言[b-Virtanen]的不同。

还有一种可能是使用连续量表。

由于连续数据通常舍入到一定的精度，为简化数据采集，可以使用如图B.3所示的评分量表。标签仅在端点使用，而标记在量表中间表示。这能够降低由于标签解释而产生的偏差。每个区域都对应一个具体的数值，且可以无歧义地采集数据。



图B.3 – 质量评分的准连续量表

B.2 额外的等级维度

如果在测试中评估的系统被认为在总体质量上基本相等，因此获得非常相似的分数，则在为每种情况准备的单独量表上评价额外的质量构成可能是有利的。通过这种方法，可以接收特定特性的信息，即在测试中目标被明显感知是不同的，尽管事实上总体质量基本是相同的。这种额外的测试结果能给出关于被测系统的极具价值的鉴别信息。

为感知整体图像质量而假定的因素的评分维度示例如下，同时还给出了该因素给质量带来的影响是积极的还是消极的：

– 亮度（积极）；
– 对比度（积极）；
– 彩色复制（积极）；
– 轮廓定义（积极）；
– 背景稳定度（积极）；
– 图像重组的速度（积极）；
– 抖动（消极）；
– “拖尾”效应（消极）；
– “蚊式”效应（消极）；
– 重影/阴影（消极）；
– 光晕（消极）。

近期研究表明，给每个因素分配适当的权重，然后将它们加在一起[b-RACE]，这些因素可以组合成感知的整体质量。

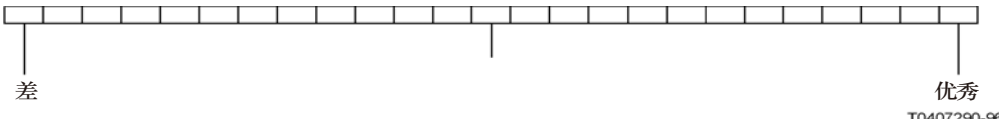
为了单独评估整体视频质量的维度，可以使用一种特定的调查问卷。在每个测试条件呈现后可提问问题的示例在下面的调查问卷中给出。

调查问卷

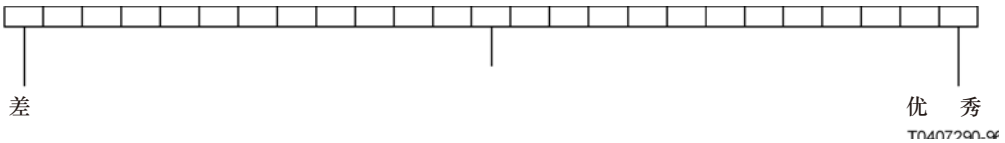
您可以回答以下关于最后一个序列的问题吗？

您可以通过在下面的量表上插入标记来表达您的意见。

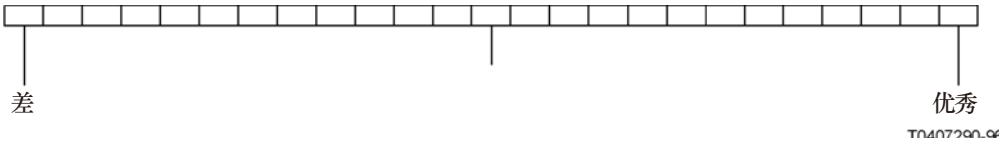
1) 您如何评价图像颜色？



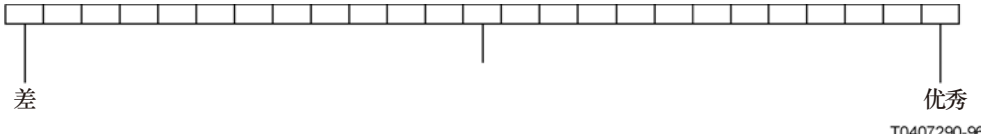
2) 您如何评价图像对比度？



3) 您如何评价图像边界？

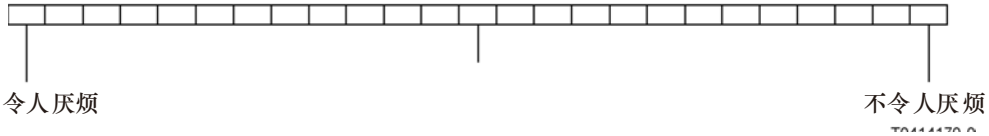


4) 您如何评价动作连续性？



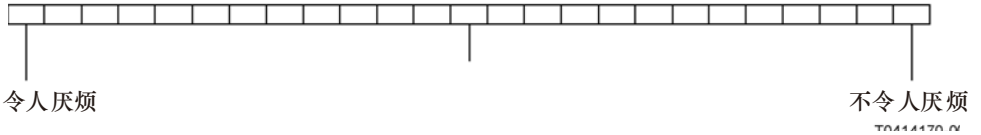
5) 您注意到序列中有任何闪烁么？ □是 □否

如果您注意到了闪烁，请在下面的量表上对其进行评价



6) 您注意到序列中有任何拖尾么？ □是 □否

如果您注意到了拖尾，请在下面的量表上对其进行评价



注 – 当使用这些量表时，涉及的所有质量/损害类别（例如，运动连续性、闪烁、拖尾等）都应在训练阶段被仔细说明。

附件C

序列对的同步呈现

（此附件是本建议书不可分割的组成部分）

C.1 序言

当被测的系统使用压缩的图片格式（例如CIF、QCIF、SIF等），且使用DCR或者PC方法时，在同一个显示器上同时显示每对的两个序列是有利的。

使用同步呈现（SP）的优点如下：

1) SP大大减少了测试的持续时间。

2) 如果使用合适的图片尺寸，被试者评估刺激间的差异会更容易。

3) 由于在相同的测试条件下，呈现的数量减半，使用SP时，被试者的关注度通常更高。

SP要求采取特别的预防措施，以避免被试者产生由于演示的类型而带来的偏差。

C.2 同步

两个序列必须被完全同步，这意味着两个序列需要在相同的帧开始和结束，同时显示也必须同步。这不排除可以比较以不同比特率编码的序列，条件是应用合适的时间上采样。

C.3 观测条件

序列必须在50％灰色背景下并排放置的两个窗口中显示（灰色在5.1节中明确），如图C.1所示。为了减少切换两个窗口的眼睛移动带来的注意力转换，观测距离应为8H，其中H表示图片高度。显示器的对角线尺寸应至少为14英寸。

图C.1 – SP中两个序列的相对位置

C.4 呈现

在DCR中，参考应被放置在同一侧（例如，左侧），被试者应了解参考和测试条件的相对位置。

在PC中，所有序列对应该以所有可能的顺序（例如，AB、BA）显示。这意味着显示在左侧的序列现在显示在右侧，反之亦然。

附件D

视频类及其属性

（此附件是本建议书不可分割的组成部分）

在本建议书中，最高视频质量是[ITU-R BT.601]建议书中定义的，4：2：2、Y、C_R、C_B格式的8位/像素线性PCM编码视频。

表D.1 – 视频类的定义

TV 0	低损耗：[ITU-R BT.601]，8位/像素，视频用于没有压缩的应用程序。
TV 1	用于完整的后期制作，有许多编辑和处理层，在设备内传输。也用于远程站点传输。与TV 0相比，感觉是透明的。
TV 2	用于简单的修改、很少的编辑、字符/标志叠加、程序插入和设施间传输。广播示例将是网络到分支的传输。其他示例是到本地前端和高质量视频会议系统的有线系统区域下行链路。与TV 0相比，感觉几乎是透明的。
TV 3	用于传输到家庭/消费者（未发生变化）。其他的示例是从本地前端到家庭和中高质量视频会议系统的有线系统。与TV 2相比，存在较低的伪影。
MM 4	所有帧编码。相对于TV 3，存在较低的伪影。中等质量视频会议。通常 ≥ 30 fps。
MM 5	可能会在编码器丢失帧。可能有能感知的伪影，但质量水平对设计任务有用，例如，低质量视频会议。
MM 6	系列静止。不旨在提供完整运动（示例：监视、图形）。

表D.2 – 视频类的属性

视频类	空间格式	传送帧速率（注1）	典型的时延延迟变化（注2）	标称视频比特率 (Mbit/s)
TV 0	[ITU-R BT.601]	最大FR	（注2）	270
TV 1	[ITU-R BT.601]	最大FR	（注2）	18至50
TV 2	[ITU-R BT.601]	最大FR	（注2）	10至25
TV 3	[ITU-R BT.601]	特殊的最大FR 帧重复	（注2）	1.5至8
MM 4a	[ITU-R BT.601]	~30或~25 fps	延迟<≈150 ms 变化<≈50 ms	~1.5
MM 4b	CIF	~30或~25 fps	延迟<≈150 ms 变化<≈50 ms	~0.7
MM 5a	CIF	10-30 fps	延迟<≈1000 ms 变化<≈500 ms	~0.2
MM 5b	≤CIF	1-15 fps	延迟<≈1000 ms 变化<≈500 ms	~0.05
MM 6	CIF-16CIF	极限→0 fps	无限制	< 0.05，极限→0 fps

注1 – 通常情况，30 fps用于525系统，25 fps用于625系统。

注2 – 广播系统都有恒定的、不一定低的单向时延和恒定延迟变化。大多数广播应用时延较低，高质量视频会议的时延介于50和500 ms间；会话类型的应用程序时延通常小于150 ms（参见[b-ITU-T G.114]建议书）。允许给定范围内的延迟变化，但这些变化不应导致感觉上令人不安的时间规整效应。

附录I

测试序列

（此附录非本建议书不可分割的组成部分）

选择合适的测试序列是规划主观性评价的关键点。当不同观测者组或不同实验室进行的测试结果必须相关时，重要的是使用一组共同的测试序列。

表I.1描述了第一组这样的序列。表中给出了每个序列的下述信息：

– 类别（表A.1中定义）；

– 关于场景的简单描述；

– 源格式（625或525行，[ITU-R BT.601-4]格式或者Betacam SP）；

– 空间和时间信息值（5.3.1和5.3.2小节分别定义）。

表I.1所列的所有序列都属公共领域，并可以自由地用于评估和示范。一些建议的序列属于[b CCIR报告1213]中描述的CCIR文库。

CCIR文库的其他序列适用于特定应用，如基于视频存储和检索的应用。

测试序列组仍在研究中。表I.1所列的测试序列组至少可以用两种方法改进或延伸：

1) 必须包括更大范围应用的代表性序列（例如，移动视频电话、远程教室等）；

2) 每个序列的源格式都应该为[ITU-R BT.601-4]格式，且包括525和625行两个版本。

表I.1 – 多媒体应用中用于视频质量评价的测试序列

序列	类别	描述	源格式	SI	TI
washdc	D	有手和铅笔运动的华盛顿特区地图	Betacam SP（525行）	130.5	17.0
3inrow	C	桌子前的男士，照相机平移	Betacam SP（525行）	81.7	30.8
vtc1nw	A	女士坐着读新闻	Betacam SP（525行）	56.2	5.3
Susie	A	年轻女士在打电话	ITU-R BT.601-4 525/625行	58.7	24.6
花园	E	景观、照相机平移	ITU-R BT.601-4 525/625行	227.0	46.4
smity2	B	售货员在有杂志的桌子前	Betacam SP（525行）	154.5	35.1

附录II

观测测试说明

（此附录并非本建议书不可分割的组成部分）

以下可用作采用ACR、ACR-HR、DCR或PC方法的实验的评估者的说明。

此外，说明应给出大概的测试持续时间、停顿、初步试验的信息以及其他对于评估者而言有用的信息。这里未包括这些信息，因为其取决于具体操作。

II.1 ACR和ACR-HR

早上好，感谢您的参与。

在本次实验中，您将会看到您正前方的屏幕上出现的简短的视频序列。序列每次出现时，您应该通过使用以下五种量级来判断其质量。

5 优秀

4 良好

3 普通

2 较差

1 差

做出判断前请仔细观察整个视频序列。

II.2 DCR

早上好，感谢您的参与。

在本次实验中，您将会看到您正前方的屏幕上出现的简短的视频序列。每个序列将快速地连续呈现两次：每对中只有第二个序列是经过处理的。在每对序列呈现结束时，您应该评估相对于第一个序列的第二个序列的损害程度。您将通过下述量级来表达您的判断：

5 无感知的

4 可感知但不令人厌烦

3 轻微令人厌烦

2 令人厌烦

1 非常令人厌烦

做出判断前请仔细观察整个视频序列对。

II.3 PC

早上好，感谢您的参与。

在本次实验中，您将会看到您正前方的屏幕上出现的简短的视频序列。每个序列将快速地连续呈现两次：每次通过不同的编解码器。序列的顺序和编解码器在序列对中的组合随机变化。在每对序列呈现结束时，您应该通过勾选如下所示的方框表达您的偏好。如果您倾向第一个序列，就勾选方框1，如果您倾向第二个序列，就勾选方框2。

做出判断前请仔细观察整个视频序列对。

附录III

持续评估的同步双重刺激

（此附录并非本建议书不可分割的组成部分）

用于持续评估的同步双重刺激（SDSCE）适用于评估稀疏性损害的影响，诸如传输错误，或者视觉信息的保真度。该方法源自[ITU-R BT.500-9]建议书中描述的单刺激连续质量评估（SSCQE）方法。

III.1 测试程序

被试者小组同时观看两个序列：一个是参考，另一个是测试条件。如果这两个序列采用标准中间格式（SIF）或更短的格式，则这两个序列可以并排在显示器上显示，不然就用两个对齐的显示器。

请被试者检查两个序列之间的差别，并通过移动手持评分设备上的滑块来判断视频信息的保真度。如果保真度理想，则滑块应放在量表范围的顶部（代码为100）；如果保真度全无，则滑块应移动到量表的底部（代码为0）。

在整个观看期间，要让被试者知道哪个序列是参考，并请他们在观看序列期间给出评分意见。

III.2 训练阶段

训练阶段是这种测试方法的一个关键部分，因为被试者可能会误解其任务。应提供书面须说明，确保所有被试者获得完全一样的信息。说明中应解释被试者将要观看的是什么，要评价的是什么（例如质量差别），以及如何表达其评分意见。被试者提出的任何问题都应得到解答，以尽可能避免因测试管理员而产生的评分偏差。

在了解说明后，应进入示范阶段。这种方式可让被试者熟悉评分程序和损害种类。

最后运行一个模拟测试，显示若干有代表性的条件。这些序列与测试中所用的序列应有所不同，应一个接一个地显示，中间没有间隔。

在模拟测试结束之后，实验者应主要检查在测试条件等同于参考序列的情况下，评价结果是否接近一百；如果情况相反，则实验者应再次进行解释和模拟测试。

III.3 测试协议功能

下述定义适用于对测试协议的说明：

- **视频段**（VS）：一个视频段对应着一个视频序列。
- **测试条件（TC）**：一个测试条件要么是一个具体的视频过程，要么是一个传输条件，也可以是二者。每个视频段（VS）应按照至少一个测试条件处理。另外，应在测试条件清单中加入参考序列，以便能够对参考/参考对进行评价。

- **阶段（S）：**一个阶段由一系列不同的成对视频段（VS）/测试条件（TC）组成，中间没有间隔，按随即顺序排列。每一阶段至少有一次含有全部VS和TC，但不必含有全部的VS/TC组合。必须由同样数目的观测者（但不一定是同样的观测者）对VS/TC的所有组合进行评分。
- **测试演示（TP）：**一个测试演示由一系列涵盖所有视频段（VS）/测试条件（TC）组合的阶段组成。
- **评分期：**请每位观测者在一测试阶段内连续评分。

III.4 数据处理

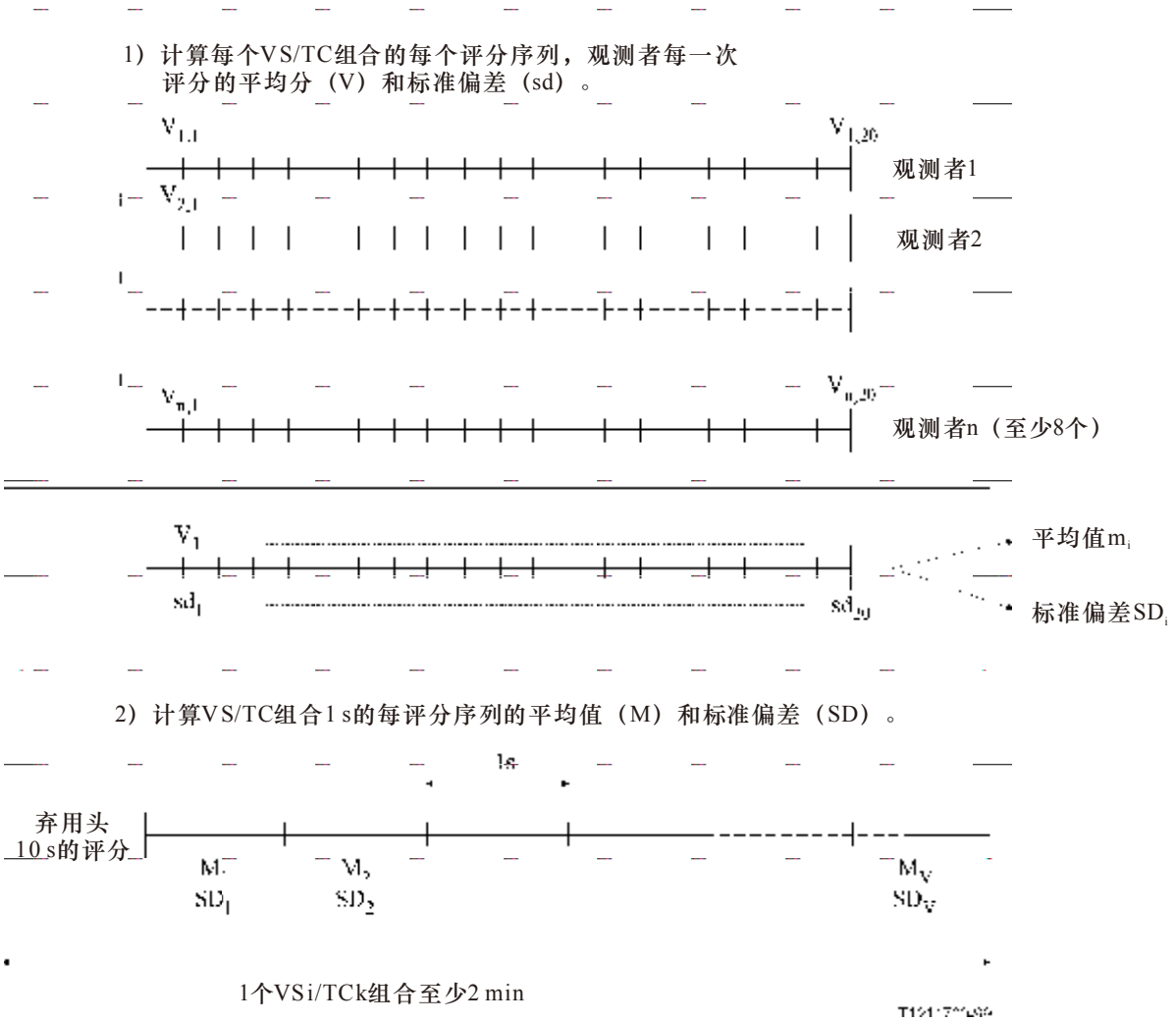
一旦测试完成，就会得到一个（或多个）数据文档，纳入了不同阶段（S）的所有评分，这些不同阶段代表了测试演示（TP）的打分总次数。通过验证每一VS/TC对都已得到处理且每一对都分配了相同次数的评分，就完成了数据有效性的第一次校验。

在按照这一协议完成的测试中收集到的数据可用3种不同的方式处理：

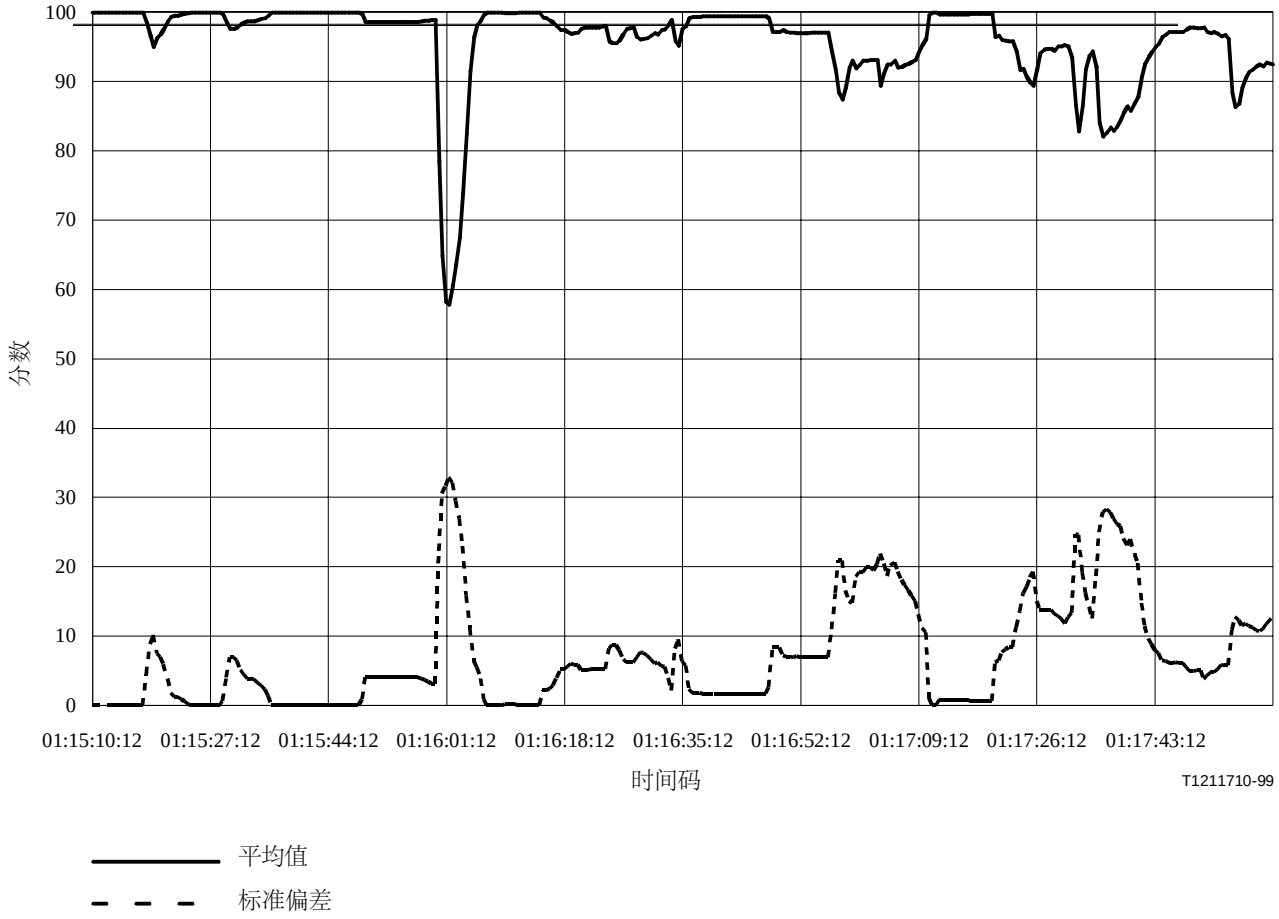
- 每一单独VS的统计分析；
- 每一单独TC的统计分析；
- 所有VS/TC对的总体统计分析。

每种情况都需要进行多步骤分析：

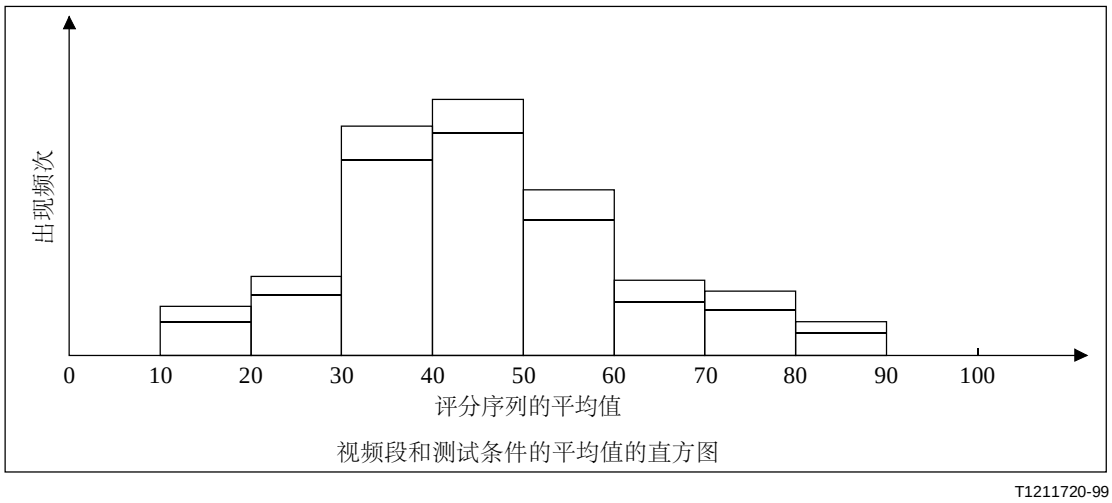
- 根据观察者的累计计算出每次评分的均值和标准差，如图III.1所示。
- 然后，每个VS被认为是最长持续时间为10 s的评分段的集合。由于近因和宽恕效应都不会影响持续不超过10 s的序列的评估，因此对于每个评分段上一步计算得出的平均值的平均和标准偏差，如图III.1所示。当需要质量变化的详细信息时，评分段的持续时间应该缩短（约为1 s）。这一步的结果可用一幅时间图表示，见图III.2。
- 分析前一步算出的均值（即与每一评分段相对应）的统计分布及其出现频次。为了避免由前一个VS × TC组合产生的近因效应，每一VS × TC样本的头10 s评分要弃用。图III.3给出了一个示例。
- 根据对出现频次的累计算出总体厌烦特性。这一计算要考虑置信区间，如图III.4所示。总体厌烦特性因为展示了每一评分段的均值与其累积出现频次之间的关系而与这一累积统计分布函数形成对应。


1）计算每个VS/TC组合的每个评分序列，观测者每一次
评分的平均分（V）和标准偏差（sd）。
观测者1
观测者2
观测者n（至少8个）
平均值mi
标准偏差SDi
2）计算VS/TC组合1 s的每评分序列的平均值（M）和标准偏差（SD）。
1s
弃用头
10 s的评分
1个VSi/TCk组合至少2 min

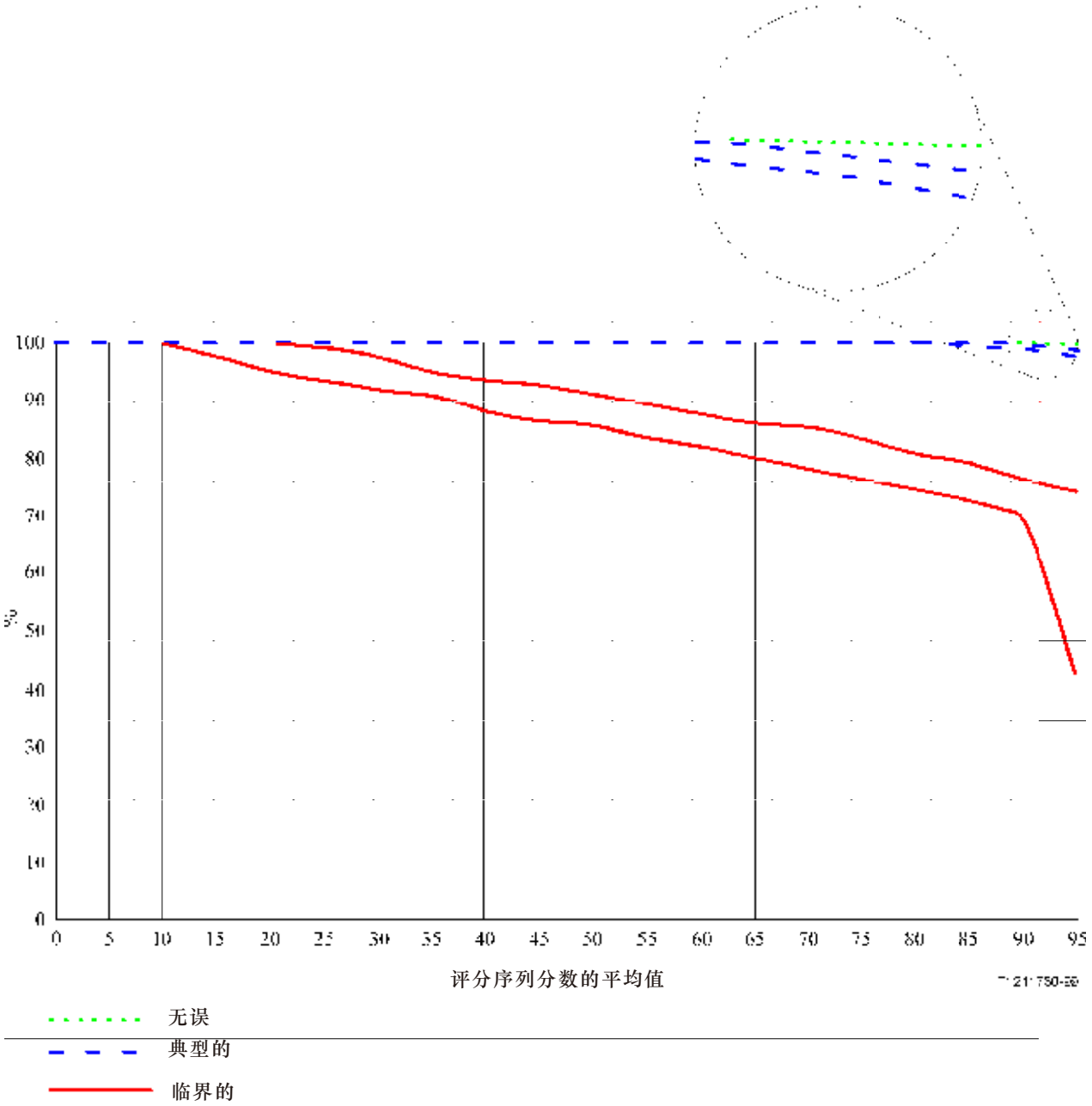
图III.1 – 数据处理



图III.2 – 原始时间图



图III.3 – 损害特性与出现频次之间的关系



图III.4 – 从统计分布中计算的、包括置信区间的总体厌烦特性

III.5 被试者信度

通过检验被试者在显示参考/参考对时的表现就可以定性评价被试者信度。在这种情况下，预计被试者将给出特别接近100的评价结果。由此证明他们了解自己要承担的任务，不会随意打分。

此外，对于SSCQE方法，可以采用与[ITU-R BT.500-9]建议书中所述程序接近的程序来检查被试者信度。

在SDSCE程序中，评分的信度取决于下面两个参数：

系统偏差：在测试期间，有的观测者可能过于乐观或过于悲观，甚或误解了评分程序（例如评分量表的含义）。这样就可能导致某一系列评分与平均系列之间或多或少存在系统偏差，甚至完全超出平均范围。

局部反演：在其他一些为人所熟知的测试程序中，观测者有时可能没有特别留心观看和跟踪所显示的序列的质量。在这种情况下，总体评分曲线相对而言尚处在平均范围内，但仍可观测到局部反演。

这两种不合意的结果（反常行为和反演）是可以避免的。参与者接受训练固然重要，但采用某种工具检测并在必要时舍弃前后不一致的观察结果也是可能的。

附录IV

基于对象的评估

（此附录非本建议书不可分割的组成部分）

应在整个场景和单个对象上评估基于对象的功能。这是因为，通常来说，由独立编码对象组成的场景可以像作者所产生的作品那样被“使用”，但在某些情况下也可以被操作，并且每个单个对象可以在完全不同的情境中使用。出于此原因，在整体场景的总体质量和每个对象的组织及轮廓的质量间找到平衡是很重要的。

因此，应该在两个运行中评估基于对象的功能（对象可扩展性和基于对象的质量可伸缩性）：

全场景评估 – 这是一种针对全序列的传统测试，包括所有的视频对象。评价方法可以是ACR（见6.1节）或DCR（见6.3节），取决于比特率的范围和源序列的临界。

基于对象的评估（OBE） – 在这种测试中，灰色背景上只显示一个视频对象，且被试者被要求去评估显示的视频对象的质量/损害（根据全场景评估中使用的测试方法）。必须明确用于视频对象的比特率百分比。评估的视频对象将从与全场景评估中完全相同的编码序列中提取。

图IV.1说明了用来评估对象可扩展性的两种测试。

图IV.1 – 评估对象可扩展性测试

在基于对象的质量可扩展性的情况下，应该进行单独的测试来评估空间可扩展性和时间可扩展性，并且只应用OBE。

对于空间和时间可扩展性，应该应用OBE来评估在同一运行中，以“参考”比特率编码的视频对象和以指定的增强比特率编码的相同视频对象。

通常来说，基于对象的功能评估应考虑整体帧的质量和单个对象的质量。前边的评估应用标准方法完成，而后边的使用OBE。

为了对基于对象编码的不同系统进行比较，实验者应提前明确分配给整体质量和单个对象质量的相对权重。

在特定的情况下，使用基于任务的评估标准而不是传统的质量评价是有意义的。例如，对用于车库的远程监控系统进行评估时，应该根据车牌的可识别性评估质量可扩展性进。实验者应根据个例、测试的目标和被研究的应用种类决定任务。

最后，对象质量评估应被应用于研究单个对象的质量对于场景总体质量的影响。这种研究的结果应用于优化基于对象的编码方案。

附录V

额外的DCR评估量表

（此附录并非本建议书不可分割的组成部分）

可以使用如图V.1所示的九级数值劣化量表。在本量表中，等级8对应于劣化的感知阈值，在此劣化等级下，观察者不完全确定能感知到劣化。

9 无感知的
8
7 可感知但不令人厌烦
6
5 轻微令人厌烦
4
3 令人厌烦
2
1 非常令人厌烦

图V.1 – 九级数值劣化量表

参考书目

[b-ITU-T G.114] ITU-T G.114（2003年），单向传输时间。

[b-ITU-T H.261] ITU-T H.261建议书（1993年），*p* × *64 kbit/s*的视听业务的视频编解码器。

[b-ITU-T P.920] ITU-T P.920建议书（1996年），视听通信的交互式测试方法。

[b-ITU-T Handbook] ITU-T手册（1993年），关于通话计时的手册，国际电联，日内瓦。

[b-ITU-R BT.812] ITU-R BT.812建议书（1992年），图文电视和类似服务中字母数字和图形图像质量的主观评价。

[b-ITU-R BT.815-1] ITU-R BT.815-1建议书（1994年），用于测量显示器对比度的信号的规范。

[b-CCIR Report 1213] CCIR 1213报告（1990年），数字编解码器主观性评价的测试图片和序列，第XI卷附件，第1部分。

[b-Gonzalez] Gonzalez, R.C. and Wintz, P.(1987), *Digital Image Processing*, 2nd Edition, Addison-Wesley Publishing Co., Reading, Massachusetts.

[b-RACE] RACE Industrial Consortium Project 1018 HIVITS, WP B5, Picture Quality Measurement, 1988.

[b-Snellen] Snellen Eye Chart.

[b-Beck] *Pseudo Isochromatic Plates* (1940), engraved and printed by The Beck Engraving Co., Inc., Philadelphia and New York, United States.

[b-Kirk] Kirk, R.E.(1982), *Experimental Design – Procedures for the Behavioural Sciences*, 2nd Edition, Brooks/Cole Publishing Co., California.

[b-Virtanen] Virtanen, M.T., Gleiss, N. and Goldstein, M.(1995), *On the use of Evaluative Category Scales in Telecommunications*, Human Factors in Telecommunication Conference, Melbourne.

[b-Guilford] Guilford, P.(1954), *Psychometric methods*, McGraw-Hill, New York.

[b-ISO/IEC 11172] ISO/IEC 11172:1993, *Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s.*

ITU-T 系列建议书

A系列	ITU-T工作的组织
D系列	一般资费原则
E系列	综合网络运行、电话业务、业务运行和人为因素
F系列	非话电信业务
G系列	传输系统和媒质、数字系统和网络
H系列	视听和多媒体系统
I系列	综合业务数字网
J系列	有线网和电视、声音节目及其他多媒体信号的传输
K系列	干扰的防护
L系列	线缆的构成、装置和保护及外部设备的其他组件
M系列	电信管理，包括TMN和网络维护
N系列	维护：国际声音节目和电视传输电路
O系列	测量设备规范
P系列	**电话传输质量、电话装置、本地线路网络**
Q系列	交换和信令
R系列	电报传输
S系列	电报业务终端设备
T系列	远程信息处理业务的终端设备
U系列	电报交换
V系列	电话网上的数据通信
X系列	数据网和开放系统通信及安全
Y系列	全球信息基础设施、互联网的协议问题和下一代网络
Z系列	用于电信系统的语言和一般软件问题

瑞士印制
2017年，日内瓦