ITU-R BT.1788 建议书

对多媒体应用中视频质量的主观评估方法

（ITU-R 102/6 号研究课题）

（2007 年）

范围

数字广播系统允许提供多媒体和数据广播应用，包括视频、音频、静态图像、文本和图表。本建议书规定评估多媒体应用视频质量的非交互式主观评估方法。

国际电联无线电通信全会，

考虑到

a） 许多国家正在引入数字广播系统；

b） 利用数字广播系统，已经引入或计划引入包括视频、音频、静态图像、文本、图表等的多媒体和数据广播服务；

c） 多媒体服务将涉及广播基础设施，其特点是可能使用固定或移动接收机、固定和可变的帧速率、不同的图像格式、先进的视频编解码、丢包等；

d） 有必要规定性能要求，并验证为各项带有性能要求的服务而考虑的技术解决方案的适宜性；

e） 这种验证将主要涉及在受控条件下对视频质量的主观评估；

f） 在 ITU-R BT.500 建议书中规定的主观评估方法可以用于多媒体应用；

g） 除了在 ITU-R BT.500 建议书中规定的那些主观评估方法外，还可使用其它的主观评估方法；

h） 采用标准方法在不同实验室间实现信息交换是十分重要的；

建议

1 测试的一般方法，即在附件 1 中所述的图像质量评估的评定尺度和观测条件，应用于实验室实验，并且只要可能，可用于多媒体应用中的运营评估；

2 在所有的测试报告中都应对测试配置、测试材料、观测者和方法做全面描述；

3 为了促进不同实验室之间的信息交换，应依据附件 2 中详述的统计方法对收集到的数据进行处理。

注 1 — 适于多媒体应用中视频质量主观评估的视频材料库，需由无线电通信第 6 研究组做进一步完善。

附 件 1

评估方法描述

1 引言

许多国家已着手部署数字广播系统，它将允许提供包括视频、音频、静态图像、文本和图表等在内的多媒体和数据广播应用。

需要标准的主观评估方法来规定性能要求，并验证为各项应用而考虑的技术解决方案的适宜性。主观方法是必要的，原因是它们提供了测量法，允许业界更直接地预测最终用户的反应。

广播系统需要交付明显不同于当前在用的多媒体应用：信息通过固定与/或移动接收机访问；帧速率可以是固定的，或者是可变的；可能的图像尺寸变化范围很大（即从 SQCIF 到 HDTV）；典型地，视频与嵌入的音频、文本与/或语音相关；视频可以通过先进的视频编解码器来处理；并且理想的观测距离很大程度上取决于应用。

在 ITU-R BT.500 建议书中规定的主观评估方法应在这一新的背景下应用。此外，可以采用新的方法完成对多媒体系统的调查，以满足用户对多媒体领域特性的要求。

本建议书描述评估多媒体应用视频质量的非交互式主观评估方法。这些方法可用于不同目的，包括但不限于：算法的选择、对视听系统性能的评定，以及在视听连接期间对视频质量等级进行评估。

与本建议书相关的术语和定义请参见附件 1 的附录 3。

2 共性

2.1 观测条件

表 1 列出了建议的观测条件。所用的显示器尺寸和类型应符合正在调查的应用。由于多媒体应用中使用了若干种显示技术，因此，所有有关评估中所用显示器的相关信息（例如制造商、型号和规范），都应予以报告。

当使用基于个人电脑的系统来展示序列时，还应报告系统的特性（例如视频显示卡）。

表 2 显示了一个有关正在测试的多媒体系统配置数据记录的例子。

如果通过使用特定的解码器－播放器组合来获取测试图像，那么这些图像必须独立于特有的外观，以便获得匿名的显示器。有必要确保质量评估不受原始环境知识的影响。

当测试中评估的系统使用降低的图形格式时，例如 CIF、SIF 或 QCIF 等，应在显示屏的一个窗口上显示片段。屏幕上背景的颜色应为 50%的灰色。

表 1

用在多媒体质量评估中的、建议的观测条件

参　　数	设　　置
观测距离(1)	限制的：1-8 H 非限制的：取决于观测者的喜好
屏幕最高亮度	70-250 cd/m^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
图形监视器背景亮度与图形亮度峰值之比(2)	≤ 0.2
背景色度(3)	D_{65}
屋内背景亮度(2)	≤ 20 lux

(1) 观测距离通常取决于应用。

(2) 该值表示允许最大可察觉失真的设置，对某些应用，允许更高值或者取决于应用。

(3) 对 PC 监视器，背景色度应尽可能接近显示器的“白点”色度。

表 2

测试中的多媒体系统的配置

参　　数	规　　范
显示器类型	
显示器尺寸	
视频显示卡	
制造商	
型号	
图像信息	

2.2 源信号

源信号直接提供基准图形以及测试中的系统的输入。源片段的质量应尽可能高。作为一个指导原则，视频信号应使用 YUV（4：2：2、4：4：4 格式）或 RGB（24 或 32 位）记录于多媒体文件中。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片段，以消除更大的变化源。

2.3 测试材料的选择

测试场景的数目和类型对解释主观评估的结果而言是至关重要的。某些过程可能导致大多数片段相同程度的损伤。在这种情况下，用少量片段（例如两个）获得的结果应提供一个有意义的评价。不过，新的系统常常具有一定的影响，这很大程度上取决于场景或片段内容。在这种情况下，应选定测试场景的数目和类型，以便为标准的节目编排提供合理的概括。此外，应为测试中的系统选定“关键但不太过度”的材料。“不太过度”这个短语指的是，场景可以仍是标准电视节目编排内容可想象的组成部分。有关场景复杂度的一个有用提示可由其空间和时间感知特性来提供。在附件 1 的附录 1 中，对空间和时间感知特性的测量有更详细的陈述。

2.4 条件和锚定的范围

由于大多数评估方法对范围变化和观测条件分布是敏感的，因此判断会议应包括变化因素的全部范围。不过，这可能与更加严格的范围近似，通过提出某些可能成为尺度极限的条件。这些可以作为例子而陈述，并确定为最大的极限（直接锚定）或分布于整个会议中，并且不被确定为最大的极限（间接锚定）。可能的话，应使用大的质量范围。

2.5 观测者

筛选后的观测者数目应至少为 15。他们应当不是专家，在某种意义上，他们与图形质量没有直接利害关系，只是作为其日常工作的一部分，并且他们不是经验丰富的评估者。在会议召开前，应使用斯内伦（Snellen）或朗多（Landolt）视力表，对观测者进行（校正）标准视觉灵敏度筛选，并使用特别选择的视力表（如 Ishihara），进行标准颜色视觉筛选。

需要的评估者数目依采用的测试程序的敏感度和可靠性而定，并取决于所追求效果的期望大小。

实验者应尽可能详细地包括其评估小组成员的特点，以利于对该因素做进一步研究。提供的建议数据可以包括：职业类别（例如广播机构职员、大学学生、办公室工作人员）、性别和年龄范围。

2.6 评估说明

应仔细向评估者介绍评估方法、损伤类型或可能出现的质量因子、等级评定尺度、时间安排等。除了那些在测试中使用、但具备可比灵敏度的训练片段外，展示待评估损伤范围和类型的训练片段应与场景一同使用。

2.7 实验设计

实验者接下来要选择实验的设计方法，以便实现特定的成本和精度目标。最好是在实验中至少包括两份复制品（即相同条件下的重复试验）。重复使计算个体的可靠性变得可能，而且如果必要，从某些对象中放弃不可靠的结果。此外，重复确保测试中的学习效果在某种程度上能够得以平衡。通过在各次测试会议开始之时包括一些“虚拟陈述”，可以在处理学习效果过程中获得进一步的改进。这些条件应是有代表性的陈述，在会议的后期予以显示。在对测试结果进行统计分析过程中，不考虑初步的陈述。

会议是一系列的陈述，不应超过半个小时。

当测试多个场景或算法时，场景或算法的陈述次序应是随机的。可能要对随机的次序进行修改，以便确保相同场景或相同算法不会出现在紧邻的时间段中（即连续地出现）。

3 评估方法

利用 ITU-R BT.500 建议书中的方法，可以对多媒体系统的视频性能进行检测。§ 3.1 提供了选定方法的列表。

§ 3.2 描述了另一种方法，称为 SAMVIQ，它利用了多媒体领域的特性，并可用于多媒体系统的性能评估。

3.1 ITU-R BT.500 建议书中的方法

以下 ITU-R BT.500 建议书中的方法，应用于评估多媒体系统的视频质量。

— 如 ITU-R BT.500 建议书附件 1 § 4 中所述的双刺激损伤尺度（DSIS）方法。

— 如 ITU-R BT.500 建议书附件 1 § 5 中所述的双刺激连续质量尺度（DSCQS）方法。

— 如 ITU-R BT.500 建议书附件 1 § 6.1 中所述的单刺激（SS）方法。

— 如 ITU-R BT.500 建议书附件 1 § 6.2 中所述的刺激－比较（SC）方法。

— 如 ITU-R BT.500 建议书附件 1 § 6.3 中所述的单刺激连续质量评估（SSCQE）方法。

3.2 多媒体视频质量（SAMVIQ）的主观评估

在该方法中，观测者准许使用一个片段的若干个版本。当所有版本都经观测者评定后，可对之后的片段内容进行评估。

不同版本可由观测者通过计算机图形接口随机选择。根据需要，观测者可以停止、评审并修改某个片段各个版本的评分。该方法包括一个显性基准（即未经处理的）片段，以及相同片段的若干个版本，这些版本包括经处理的和未经处理的（即隐含基准）片段。片段的各个版本都单独显示，并使用一个类似于在 DSCQS 方法中使用的连续质量尺度来评价。因此，该方法在功能上与利用随机访问的单刺激方法十分类似，但只要观测者想要观测，他就可以观测显性基准，这使得该方法类似于使用一个基准的方法。

SAMVIQ 质量评估方法使用连续质量尺度，以提供对视频片段内在质量的测量。各个观测者在从 0 到 100 评级的连续尺度上移动一个滑条，该连续尺度用 5 个线性排列的质量项目来注释（很好、好、一般、差、很差）。

逐个场景地进行质量评估（见图 1），包括显性基准、隐含基准和各种各样的算法。

为更好地理解这一方法，定义了以下特定词汇：

场景：视听内容；

片段：综合处理过或未经处理的场景；

算法：一种或多种图像处理方法。

3.2.1 显性、隐含的基准与算法

评估方法通常包括质量锚，以稳定结果。在 SAMVIQ 方法中，出于以下原因，考虑了两个高质量锚。已经完成的一些测试表明，可以使用*显性基准*来最大限度地缩小分值的标准差，而不使用隐含的基准或不使用基准。尤其是对多媒体数字信号编解码器性能的评估，最好使用显性基准来获得最可靠的结果。为了评估基准的内在质量，也可加上*隐含基准*，而不是显性参考，原因是陈述是匿名的，并且是经过处理的片段。显性名称“基准”会对大约 30%的观测者产生影响。这些观测者对显性基准可能给出最高分（100 分），而该分值总的说来有别于隐含基准对应的分值。值得注意的是，当没有可用的基准时，测试仍有可能进行，但标准的偏差会显著增大。

SAMVIQ 方法适用于多媒体内容，原因是它可能结合图像处理的不同特点，例如多媒体数字信号编解码器类型、图像格式、比特率、时间更新、图像缩放等。*算法*这个名称总结了这些特点的其中一个特点或其组合。

3.2.2 测试条件

在场景期间，临界点的变化是受到限制的，原因是在其它方法（提供一个综合分值，如单刺激方法）隐含使用的相同规则后选择同样的内容。最大的片段观测期为 10 秒或 15 秒，对获得稳定的和可靠的质量分值而言，这已足够。应使用专用的解码器－播放器或其产品的屏幕拷贝，以保持适当的显示性能。

3.2.3 测试机构

a） 如图 1 所示，逐个场景地进行测试。

b） 对当前场景，可能以任何次序来播放任何片段，并为其打分。每个片段都可以多次播放和打分。

c） 从一个场景到另一个场景，对片段的访问是随机的，防止观测者试图根据已排好的次序、以完全相同的方式来做出判定。实际上，在一个测试中，算法的次序仍保持相同，以便简化对结果的分析和陈述。只有来自相同按钮的相应访问是随机的。

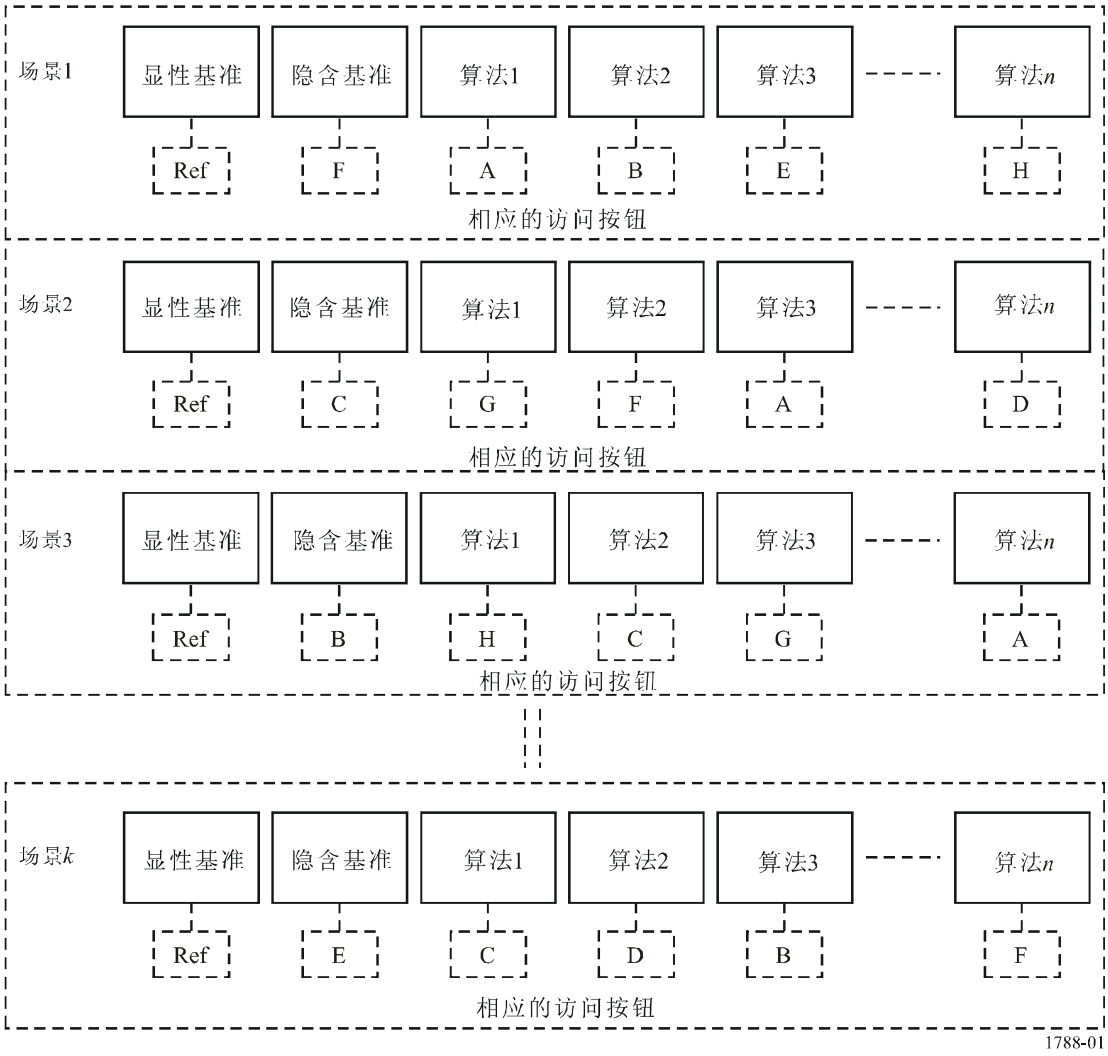
d） 对第一次观测，当前的片段必须在打分之前全部播放过；否则，可能立即打分和停止。

e） 为测试下一个场景，必须为当前场景的所有片段打分。

f） 为完成测试，必须为所有场景的所有片段打分。

图1

SAMVIQ方法的测试机构举例



SAMVIQ 方法通过软件来实现。除了图 1 中所示的访问按钮，“播放”、“停止”、“下一个场景”和“上一个场景”按钮都是必需的，以便允许观测者管理不同场景的表述（例如，参见附件 1 的附录 2）。当观测者已给出一个分值，那么应在该场景对应的访问按钮下方显示出来。当一个片段的所有不同版本都已经过评级时，仍允许观测者对分值进行比较，并且如有必要，可以对分值进行修改。不必评估当前的整个片段，原因是，在第一遍观测中，已经突出了大的差别。

附 件 2

数据表述与分析

1 摘要信息

为了复制测试或比较不同测试的结果，需要提供有关测试环境的精确数据。因此，如表 3 所示，建议报告有关测试环境的信息。

表 3

测试摘要信息

方法名称	
显示技术	
显示器的参考名称	
最大亮度等级（cd/m²）	
黑色亮度等级（cd/m²）	
黑色等级设置：PLUGE（前面所述可察觉的黑色等级距离门限＝8）。否则表示门限值。	
背景亮度等级（cd/m²）	
亮度（lux）	
观测距离： — 不受限制的：在显示器之前 — 受限制的：nH	
显示器尺寸（对角线，以英寸表示）	
宽/高显示比	
显示格式（行与列的数目）	
图像输入格式（行与列的数目）	
图像输出格式(1)（行与列的数目）	
白色色温：D65 否则 白色彩色坐标（x，y）	
有效观测者数目	

(1) 当处理输入图像时，例如在显示器上重新调节输入图像时，需要该信息。

显示特性可能影响测试结果。其它信息，例如亮度响应（百万分之一的保真度）和基色，应对平板显示器提出要求。

视频片段的特性对设计测试或解释其结果而言是重要的。如附件 1 的附录 1 所述，建议报告空间－时间特性。该信息应在视频材料库中收集适于多媒体应用中视频质量主观评估的测试片段时加以考虑。

2 分析方法

分析方法是 ITU-R BT.500 建议书附件 2§ 2 中所述的那些方法。

3 观测者的筛选

对附件 1§ 3.1 中列举的方法，筛选程序如 ITU-R BT.500 建议书附件 2§ 2.3 所述。

下一节描述对 SAMVIQ 的筛选。不过，该程序可用于 SS、DSIS 和 DSCQS 方法。该程序执行起来比 ITU-R BT.500 建议书中有关那些方法的相应的筛选程序更简单。

3.1 SAMVIQ 筛选程序

各观测者必须用稳定的和相关的方法来对各个场景和算法的质量大幅下降做出判断。舍弃准则依据所有观测者对某个给定测试会议的平均分来验证某个观测者分值的一致程度。与在 DSQCS 方法中一样，在 SAMVIQ 方法中，可以考虑所有算法（隐含的基准、低锚、经编码的片段）。判定准则基于测试的所有观测者相应的平均分对应的单个分值相关性。

3.2 皮尔森相关

质量尺度与观测者分值范围之间的关系被认为是线性的，以便应用皮尔森相关。

主要目的是，如果某个观测者的分值与整个测试会议所有观测者的平均分一致，那么用一种简单的方法来验证。隐含的基准被认为是高质量的锚。如果包括了低的和高的锚，那么它们提高了相关值，观测者之间的相关偏移值反而降低了。

$$r(x,y)=\frac{\sum_{i=1}^{n}x_i y_i-\frac{\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n}}{\sqrt{\left(\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n}y_i^2-\frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n}\right)}}$$

其中：

x_i： 表示三个参数（算法，比特率，场景）所有观测者的平均分；

y_i： 对同样的三个参数，某个观测者的单个分值；

n： （算法数目）×（场景数目）；

i： {多媒体数字信号编解码器数目，比特率数目，场景数目}。

3.3 斯皮尔曼等级相关

即使不认为质量尺度与观测者分值范围之间的关系是线性的[1]，也可以应用斯皮尔曼等级相关。

$$r(x,y)=\left[1-\frac{6\times\sum_{i=1}^{n}\left[R(x_i)-R(y_i)\right]^2}{n^3-n}\right]$$

其中：

x_i： 对三个参数（算法、比特率、场景）所有观测者的平均分；

y_i： 对相同的三个参数，某个观测者的单个分值；

n： （算法数目）×（场景数目）；

$R(x_i$ 或 $y_i)$： 排列次序；

i： {编解码数目、比特率数目、场景数目}。

3.4 放弃一名测试观测者的最终舍弃标准

根据以下条件，为了放弃观测者，执行斯皮尔曼等级和皮尔森相关：

IF [均值（r）－标准差（r）] > 最大相关门限（MCT）
舍弃门限＝最大相关门限（MCT）
ELSE 舍弃门限＝[均值（r）－标准差（r）]

IF [r（观测者 $_i$）] > 舍弃门限
THEN 不放弃测试的观测者“i”
ELSE 放弃测试的观测者“i”

其中：

r＝ 最小（皮尔森相关，斯皮尔曼等级相关）

均值（r）： 测试的所有观测者相关的平均值

标准差（r）：测试的所有观测者相关的标准差

最大相关门限（MCT）＝0.85。

最大相关门限值 0.85 对 SAMVIQ 和 DSCQS 方法是有效的，否则，对 SS 和 DSIS 方法必须考虑最大相关门限值 0.7。

---

[1] 通常，皮尔森相关结果非常接近斯皮尔曼相关结果。

附件 1 的附录 1

空间和时间信息测量

以下给出的空间和时间信息测量法为完整测试片段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法用最大函数（片段的最大值）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。在测试片段上使用信息分发也允许用场景剪辑对场景进行更好的评估。

空间感知信息（SI）：是一种通常用于表示图形空间细节数量的测量法。它通常高于空间上更复杂的场景。它并不意味着是一种信息熵测量法，也与通信理论中定义的信息无关。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 滤波器滤除后的帧中像素的标准差（std_{space}）。为视频片段中各个帧重复该操作，产生场景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。该过程可以用方程式的形式来表示：

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

时间感知信息（TI）：是一种通常用于表示视频片段时间变化次数的测量法。它通常高于高速运动的片段。它并不意味着是一种信息熵测量法，也与通信理论中定义的信息无关。

时间信息测量法 *TI*，当作所有 i 和 j 的空间（std_{space}）上的标准差最大时间值（$\max_{time}$）来计算。

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

其中，M_n（i，j）指的是帧中相同位置上各像素之间的差异，但属于两个随后的帧，就是说：

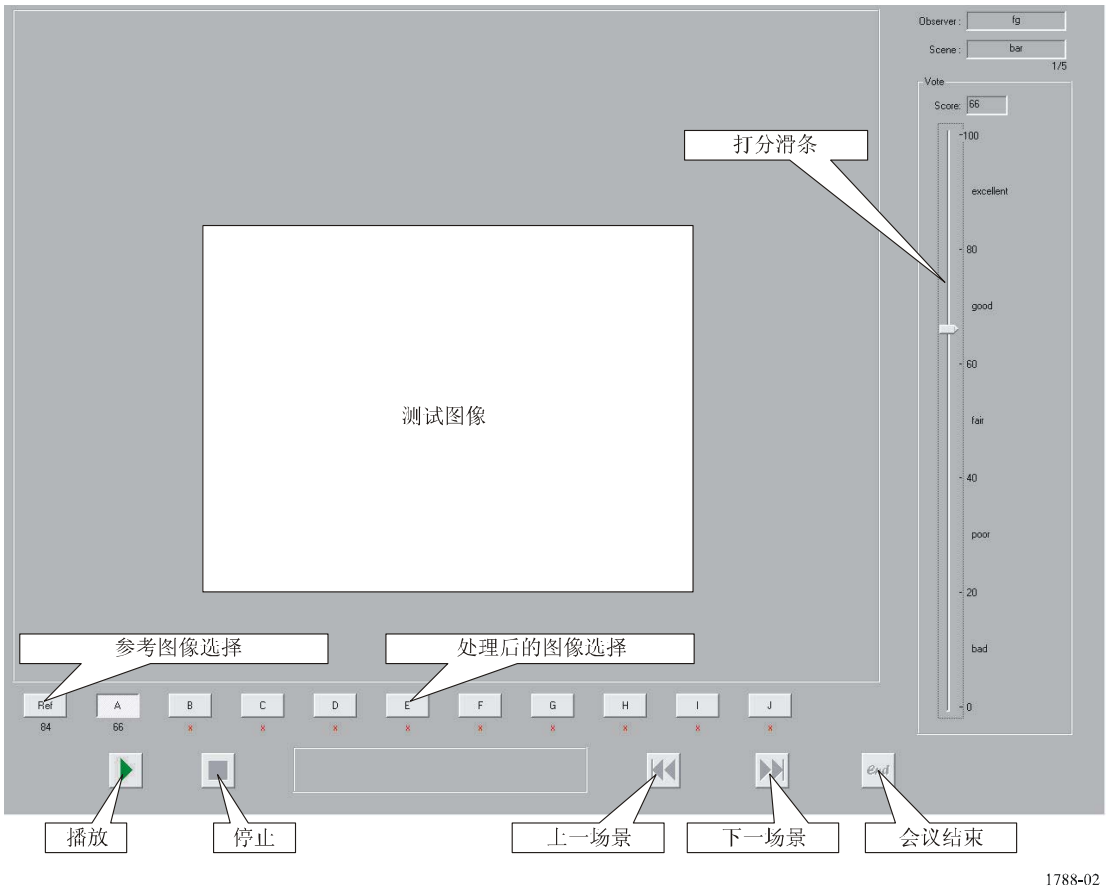
$$M_n(i, j) = F_n(i, j) - F_{n-1}(i, j)$$

其中，F_n（i，j）是时间上第 n 帧第 i 行和第 j 列处的像素。

注 1 — 对包含场景剪辑的场景，可以给定两个值：在一个值中，时间信息测量法中包括了场景剪辑，在另一个值中，测量法不包括场景剪辑。

附件 1 的附录 2

SAMVIQ 接口举例



1788-02

附件 1 的附录 3

术语和定义

Algorithm（算法）	一项或多项图像处理操作
AVI	音频视频交错
CCD	电荷耦合器件
CI	置信区间

CIF	通用中间格式（H.261 建议书中为视频电话定义的图形格式：352 行×288 像素）
CRT	阴极射线管
DSCQS	使用连续质量尺度方法的双刺激
DSIS	使用缺损尺度方法的双刺激
LCD	液晶显示器
MOS	平均评分法
SC	刺激比较方法
PDP	等离子显示板
PS	节目片段
QCIF	四分之一 CIF 格式（H.261 建议书中为视频电话定义的图形格式：176 行×144 像素）
SAMVIQ	多媒体视频质量的主观评估
Sequence（片段）	经综合处理或未经处理的场景
Scene（场景）	视听内容
S/*N*	信噪比
SI	空间信息
SIF	标准中间格式[ISO 11172（MPEG-1）中定义的图形格式：352 行×288 像素×25 帧/秒和 352 行×240 像素×30 帧/秒]
SP	同时表述
SQCIF	子 QCIF
SS	单刺激法
SSCQE	使用连续质量评估方法的单刺激
std	标准差
TI	时间信息
TP	测试表述
TS	测试会议
VTR	磁带录像机