# 一种实用的感知视频质量度量方法

作者：智丽、安妮·亚伦、伊奥尼斯·卡萨沃尼迪斯、阿努什·穆尔西和梅加·马诺哈拉

在Netflix，我们关心视频质量，我们也关心视频质量的精确测量。我们的方法，视频多方法评估融合（VMAF），旨在反映观众对我们的流媒体质量的看法。我们正在公开采购这一工具，并邀请研究界与我们在这一重要项目上进行合作。

### 我们对高质量视频的追求

我们努力为我们的会员提供一个良好的观看体验：流畅的视频播放，没有烦人的图片伪影。考虑到网络带宽和观看设备的限制，这项努力的一个重要部分是尽可能提供具有最佳感知质量的视频流。我们经过多方努力，不断朝着这个目标努力。

首先，我们在视频编码领域进行了创新。流式视频需要使用标准（如H.264/AVC、HEVC和VP9）进行压缩，以便以合理的比特率进行流式传输。当视频压缩过多或不正确时，这些技术会导致质量受损，称为压缩伪影。专家称之为“阻塞”、“响铃”或“蚊子噪音”，但对于典型的观众来说，视频只是看起来不对。为此，我们定期比较编解码器供应商的压缩效率、稳定性和性能，并集成市场上最好的解决方案。我们评估不同的视频编码标准，以确保我们保持在压缩技术的前沿。例如，我们对H.264/AVC、HEVC和VP9进行了比较，在不久的将来，我们将对AOM和JVET开发的下一代编解码器进行实验。即使在既定的标准内，我们继续实验配方决策（见）和速率分配算法，以充分利用现有的工具集。

我们在中对Netflix视频流进行编码，这使我们能够扩展以满足业务需要。为了将坏的源代码交付、软件错误和云实例的不可预测性（暂时性错误）的影响降到最低，我们在管道中的各个点自动进行质量监控。通过这种监控，我们试图检测视频质量问题在摄取和在我们的管道中的每个转换点。

最后，当我们在Netflix生态系统的各个领域迭代（如自适应流媒体或内容交付网络算法）并运行时，我们努力确保通过系统改进保持或改进视频质量。例如，旨在减少回放开始延迟或重新缓冲的自适应流算法的改进不应降低流会话中的整体视频质量。

上述所有具有挑战性的工作都取决于一个基本前提：我们能够准确有效地测量视频流的感知质量。传统上，在视频编解码器的开发和研究中，人们广泛使用两种方法来评估视频质量：1）视觉主观测试和2）简单指标（如PSNR）的计算，或者最近的SSIM[1]。

毫无疑问，人工目视检查在操作上和经济上是不可行的

对于我们生产的吞吐量，A/B测试监控和编码研究实验。图像质量的测量是一个古老的问题，对此提出了一些简单实用的解决方案。均方误差（MSE）、峰值信噪比（PSNR）和结构相似性指数（SSIM）是最初为图像设计的度量标准，后来扩展到视频。这些度量通常在编解码器（“循环中”）中用于优化编码决策和报告编码视频的最终质量。尽管该领域的研究人员和工程师都清楚，PSNR并不能始终如一地反映人类的感知，但它仍然是编码比较和编码标准化工作的事实标准。

### 构建Netflix相关数据集

为了评估视频质量评估算法，我们采用了数据驱动的方法。第一步是收集与我们的用例相关的数据集。尽管有用于设计和测试视频质量指标的公共数据库，但它们缺乏与Netflix等实际流媒体服务相关的内容多样性。就源和编码的质量而言，它们中的许多不再是最先进的；例如，它们包含标准定义（SD）内容，并且仅涵盖较旧的压缩标准。此外，由于评估视频质量的问题远比测量压缩伪影更普遍，现有的数据库试图捕获不仅由压缩引起的更大范围的损伤，而且还包括传输损失、随机噪声和几何变换。例如，实时传输典型的黑白、低分辨率视频（640×480）的监控画面显示出与在客厅观看自己喜爱的Netflix节目时明显不同的观看体验。

Netflix的流媒体服务提供了一组独特的挑战，同时也为设计一个能够准确反映流媒体视频质量的感知指标提供了机会。例如：

视频源特性。Netflix拥有大量的电影和电视节目，这些节目在类型上表现出多样性，如儿童内容、动画、快速动作电影、带有原始镜头的纪录片等。此外，它们还表现出不同的低层次源特性，如电影颗粒、传感器噪声、计算机生成的纹理、一致的黑暗场景或者非常鲜艳的颜色。过去开发的许多质量度量没有经过调整以适应源内容中的这种巨大变化。例如，许多现有的数据库缺乏动画内容，大多数不考虑电影颗粒，这是非常普遍的专业娱乐内容的信号特征。

文物的来源。由于Netflix视频流是使用健壮的传输控制协议（TCP）传输的，因此数据包丢失和比特错误永远不是视觉损害的来源。这将在编码过程中留下两种类型的伪影，它们最终将影响观看者的体验质量（QoE）：压缩伪影（由于有损压缩）和缩放伪影（对于较低的比特率，视频在压缩之前被降采样，然后在观看者的设备上被升采样）。通过裁剪一个质量度量来只覆盖压缩和缩放工件，以通用性换取精度，它的精度有望超过通用性度量。

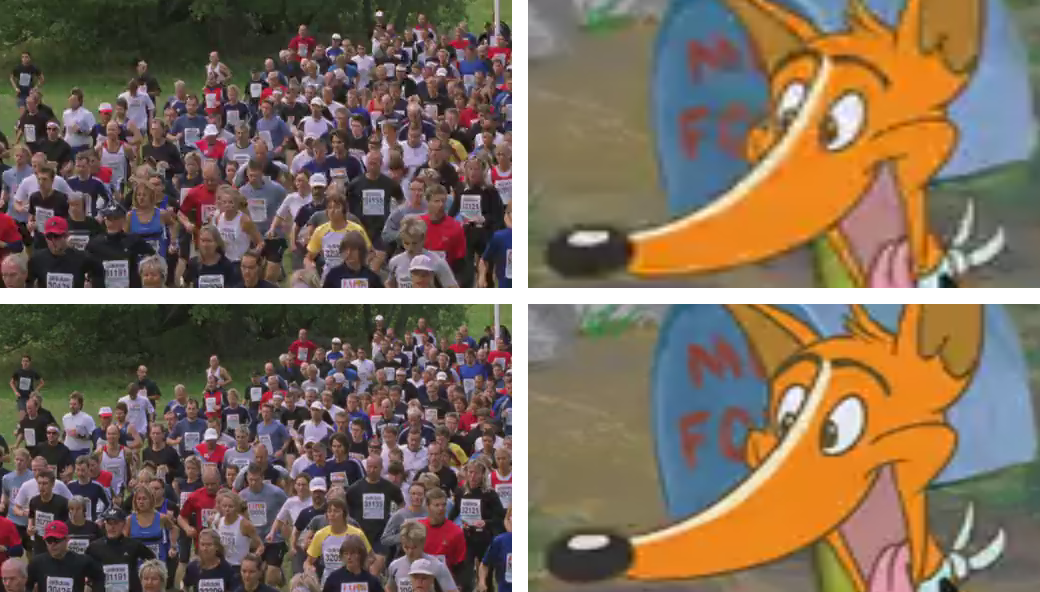
为了构建更适合Netflix用例的数据集，我们从Netflix目录中的流行电视节目和电影中选择了34个源片段（也称为参考视频），每6秒长一个样本，并将它们与一些公共可用片段组合在一起。源剪辑涵盖了广泛的高级特征（动画、室内/室外、相机运动、面部特写、人物、水、明显的显著性、对象数量）和低级特征（胶片颗粒噪声、亮度、对比度、纹理、运动、颜色变化、颜色丰富度、锐度）。使用源片段，我们以384×288到1920×1080的分辨率和375 kbps到20000 kbps的比特率对H.264/AVC视频流进行编码，得到大约300个失真视频。这将扫描广泛的视频比特率和分辨率，以反映Netflix成员广泛变化的网络状况。

然后，我们进行主观测试，以确定非专家观察员如何对编码视频相对于源剪辑的损害进行评分。在标准化主观测试中，我们使用的方法被称为双刺激损伤量表（DSIS）方法。参考和失真视频在消费者级电视上顺序显示，环境照明受控（如建议ITU-R BT.500–13[2]中规定）。如果失真视频的编码分辨率小于参考分辨率，则在将其显示在电视上之前，会将其放大到源分辨率。观察者坐在类似客厅的沙发上，被要求将损伤程度从1分（非常烦人）到5分（不明显）进行评分。将所有观察者的得分合并，为每个失真视频生成差异平均意见得分或DMOS，并在0到100的范围内进行标准化，参考视频的得分为100。本文将参考视频、扭曲视频和观察员的DMOS分数集称为NFLX视频数据集。

### 传统视频质量指标

传统的、广泛使用的视频质量指标如何与NFLX视频数据集的“基本事实”DMOS分数相关联？

#### 形象的例子



上面，我们看到从4个不同的失真视频中捕获的静止帧的部分；上面的两个视频报告的PSNR值约为31 dB，而下面的两个视频报告的PSNR值约为34 dB。然而，人们几乎无法注意到“人群”视频的差异，而在两个“福克斯”视频上的差异更为明显。人类观察者通过将这两个“人群”视频的DMOS评分分别为82分（上）和96分（下）来证实这一点，同时将这两个“福克斯”视频的DMOS评分分别为27分和58分。

#### 详细结果

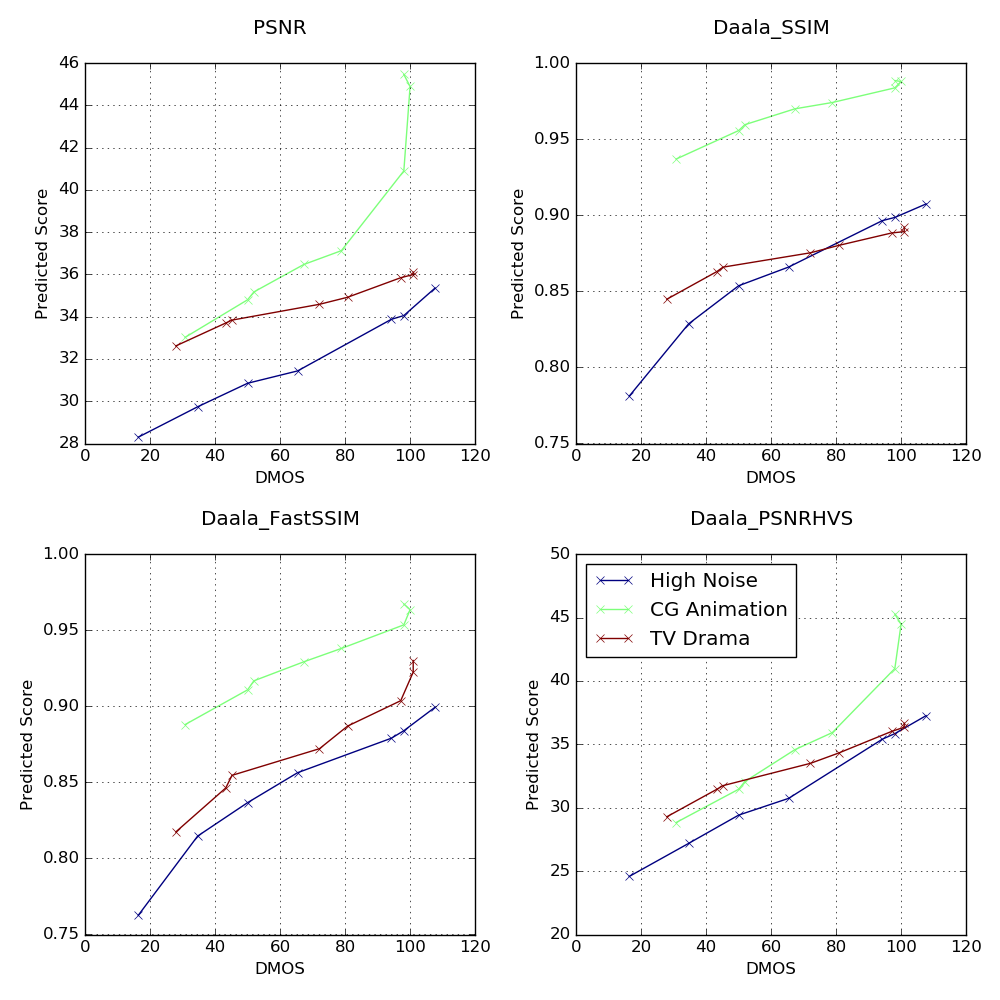
下面的图是散点图，示出了X轴上观察者的DMOS和来自Y轴上的不同质量度量的预测得分。这些图来自NFLX视频数据集的选定子集，我们将其标记为NFLX-TEST（有关详细信息，请参阅下一节）。每个点代表一个扭曲的视频。我们为四个质量指标绘制结果：

* 亮度分量
* SSIM[1]
* 多尺度FastSSIM[3]
* PSNR-HVS[4]

有关SSIM、多尺度快速SSIM和PSNR-HVS的更多详细信息，请参阅参考章节中列出的出版物。对于这三个度量，我们在代码库[5]中使用了实现，因此后续图中的标题以“Daala”作为前缀。

从图表中可以看出，这些指标未能提供一致预测观察员的决策支持系统评级的分数。例如，聚焦于左上角的PSNR图，对于35db左右的PSNR值，“地面真值”DMOS值的范围从10（损害是烦人的）到100（损害是不可察觉的）。对于SSIM和多尺度FastSSIM度量也可以得出类似的结论，其中接近0.90的分数可以对应10到100的DMOS值。如ITU-R BT.500–13[2]附件3.1所述，在每个图上方，我们报告每个指标的（SRCC）、（PCC）和均方根误差（RMSE）数字，这些数字是在非线性逻辑拟合后计算的。SRCC和PCC值接近1.0，RMSE值接近0是可取的。在四个指标中，PSNR-HVS的SRCC、PCC和RMSE值最好，但仍缺乏预测精度。

为了在各种各样的内容中实现有意义的性能，度量应该表现出良好的相对质量分数，即度量中的delta应该提供有关感知质量delta的信息。在下图中，我们选择了三个典型的参考视频，一个高噪声视频（蓝色）、一个CG动画（绿色）和一个电视剧（rust），并为每个视频绘制不同失真视频的预测分数与DMO。为了有效地作为一个相对质量分数，在相同的质量曲线范围内不同剪辑之间的恒定斜率是可取的。例如，参考下面的PSNR图，在34分贝到36分贝的范围内，电视剧的PSNR变化约2分贝对应于DMOS变化约50（50到100），而CG动画的相同范围内的类似2分贝变化对应于DMOS变化小于20（40到60）。虽然SSIM和FastSSIM在CG动画和电视剧剪辑方面表现出更为一致的斜率，但它们的性能仍然不足。



总之，我们看到传统的度量标准对我们的内容不起作用。为了解决这个问题，我们采用了一个基于机器学习的模型来设计一个指标，试图反映人类对视频质量的感知。下一节将讨论此度量。

### 我们的方法：视频多方法评估融合（VMAF）

在我们与南加州大学（University of Southern California）[6][7]的C.-C.J.Kuo教授及其团队的研究合作基础上，我们开发了视频多方法评估融合（VMAF），它通过组合多个基本质量度量来预测主观质量。基本原理是，每个基本度量在源内容特征、工件类型和失真程度方面可能有其自身的优势和劣势。通过使用机器学习算法 - 将基本度量“融合”为最终度量，在我们的例子中，支持向量机（SVM）回归器 - 为每个基本度量分配权重，最终度量可以保留单个度量的所有优点，并提供更准确的最终分数。机器学习模型是通过主观实验（在我们的例子中，是NFLX视频数据集）获得的意见得分来训练和测试的。

当前版本的VMAF算法和模型（表示为VMAF 0.3.1）作为VMAF开发工具包开源软件的一部分发布，使用了以下由支持向量机（SVM）回归融合的基本指标[8]：

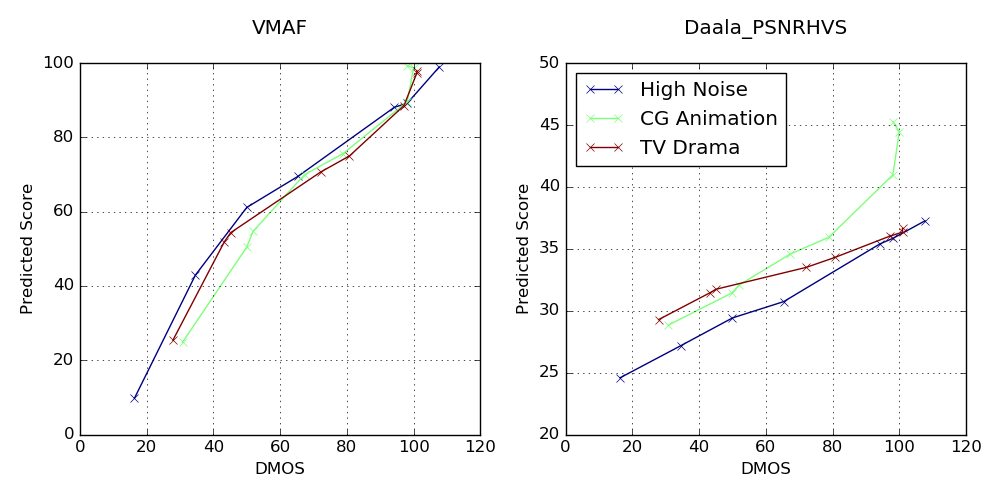
* 视觉信息保真度（VIF）[9]。VIF是一种基于质量是信息保真度损失度量的补充的图像质量度量方法。在最初的形式中，VIF分数是以四个量表相结合的保真度损失来衡量的。在VMAF中，我们采用了VIF的一个改进版本，其中每个尺度的保真度损失都包含在一个基本度量中。
* 详细损失度量（DLM）[10]。DLM是一种基于分别测量影响内容可见性的细节损失和分散观众注意力的冗余损害的原理的图像质量度量。原始指标结合了DLM和附加损伤测量（AIM）得出最终得分。在VMAF中，我们只采用DLM作为基本度量。特别注意一些特殊情况，如黑框，在这些情况下，原始公式的数值计算会出现故障。

VIF和DLM都是图像质量指标。我们进一步介绍了以下简单的特征来说明视频的时间特性：

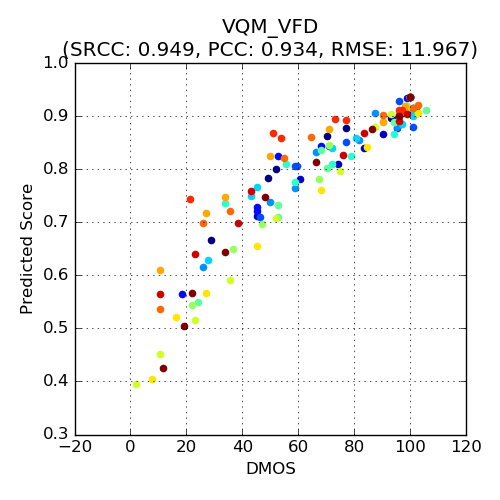
* 动议。这是测量相邻帧之间时间差的简单方法。这是通过计算亮度分量的平均绝对像素差来实现的。

这些基本的度量和特性是通过测试和验证的迭代从其他候选中选择的。

我们将VMAF的准确性与上述其他质量指标进行了比较。为了避免VMAF对数据集的不公平过度拟合，我们首先将NFLX数据集分成两个子集，称为NFLX-TRAIN和NFLX-TEST。这两组具有不重叠的参考剪辑。然后用NFLX-TRAIN数据集训练SVM回归器，并在NFLX-TEST上进行测试。下图显示了NFLX-TEST数据集和所选参考剪辑 - 高噪声视频（蓝色）、CG动画（绿色）和电视剧（rust）上VMAF指标的性能。为了便于比较，我们重复了PSNR-HVS图，这是前一节中性能最好的度量。很明显，VMAF的性能明显更好。



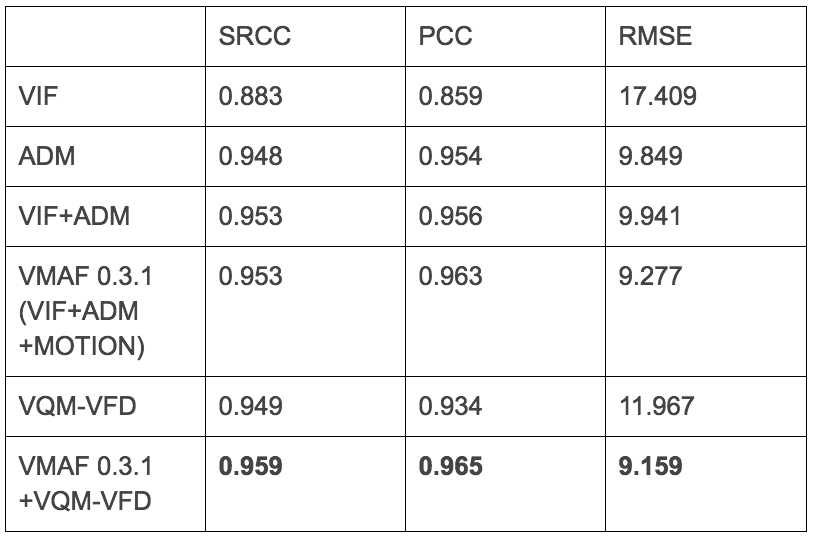
我们还将VMAF与（VQM-VFD）[11]进行了比较，许多人认为它是该领域的最新技术。VQM-VFD是一种使用神经网络模型将低级特征融合到最终度量中的算法。它在精神上类似于VMAF，只是它提取了较低层次的特征，如空间和时间梯度。



很明显，VQM-VFD在NFLX-TEST数据集上执行接近VMAF的操作。由于VMAF方法允许将新的基本度量合并到其框架中，VQM-VFD也可以作为VMAF的基本度量。

下表列出了由SRCC、PCC和RMSE图测量的VMAF模型在融合NFLX-测试数据集上的各个基本度量的不同组合后的性能，以及VMAF 0.3.1的最终性能。我们还列出了使用VQM-VFD增强VMAF的性能。这些结果证明了我们的前提，即高性能质量指标的智能融合会增加与人类感知的相关性。

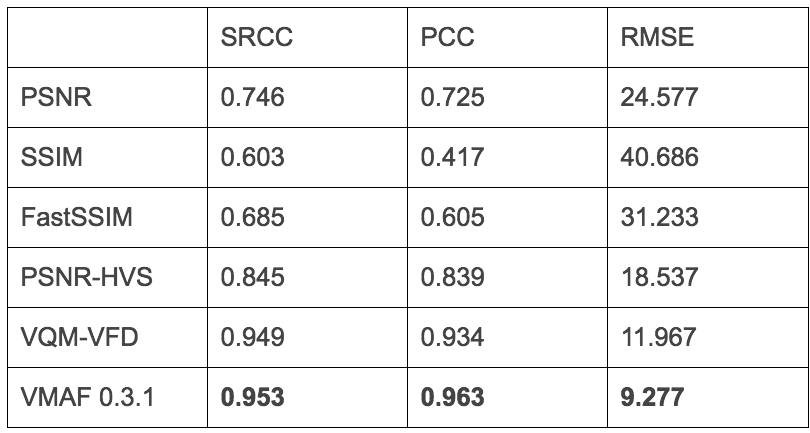
#### NFLX-测试数据集



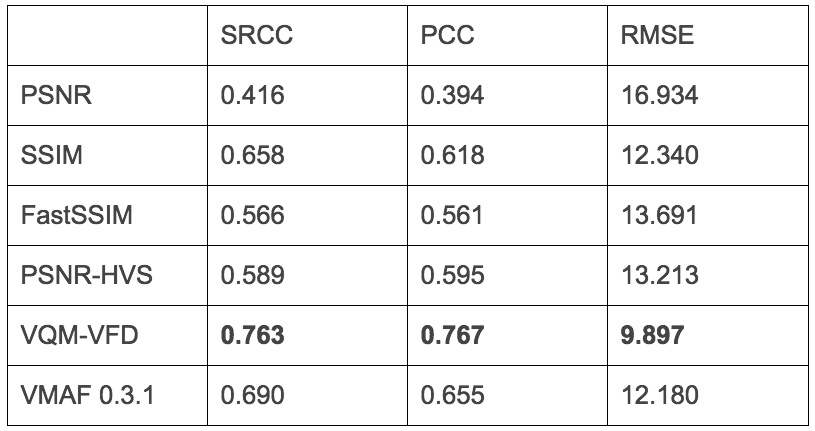
### 结果摘要

在下表中，我们总结了前面讨论的不同度量的SRCC、PCC和RMSE，这些度量是在NLFX-TEST数据集和三个流行的公共数据集上进行的：VQEG HD（仅限vqeghd3集合）[12]、实时视频数据库[13]和实时移动视频数据库[14]。结果表明，除了实时数据集外，VMAF 0.3.1的性能优于其他指标，与性能最好的VQM-VFD相比，vmaf0.3.1仍然具有竞争力。由于VQM-VFD在四个数据集之间显示出良好的相关性，我们正在试验VQM-VFD作为VMAF的基本度量；尽管它不是开源版本VMAF 0.3.1的一部分，但它可以集成到后续版本中。

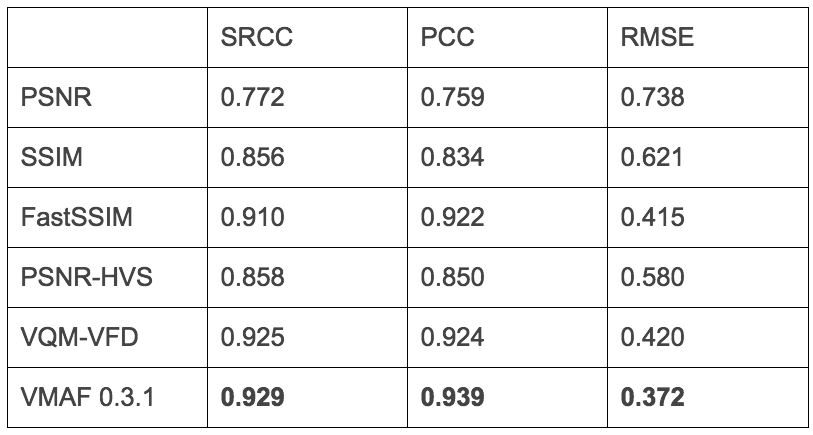
#### NFLX-测试数据集



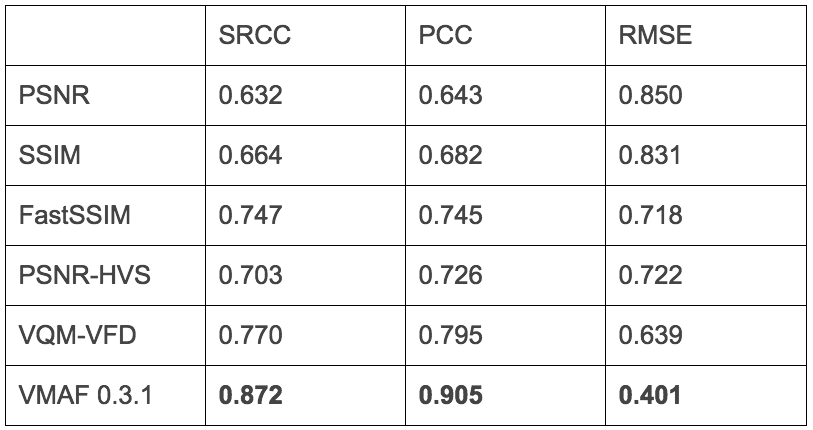
#### 实时数据集\*



#### VQEGHD3数据集\*



#### 实时移动数据集



### VMAF开发工具包（VDK）开源包

为了在互联网上提供高质量的视频，我们认为该行业需要良好的感知视频质量指标，这些指标实用且易于大规模部署。我们开发了VMAF来帮助我们解决这个需求。今天，我们将在Apache许可证版本2.0下开放源代码。通过开源VDK，我们希望它可以随着时间的推移而发展，从而提高性能。

VDK核中的特征提取（包括基本度量计算）部分是计算密集型的，因此为了提高效率，用C语言编写。控制代码是用Python编写的，用于快速原型设计。

该包附带了一个简单的命令行界面，允许用户在单模式（run\_VMAF command）或批处理模式（run\_VMAF\_in\_batch command）下运行VMAF，可以选择启用并行执行。此外，由于特征提取是最昂贵的操作，用户还可以将特征提取结果存储在数据存储中，以便以后重用。

该软件包还为进一步定制VMAF模型提供了一个框架，其基础是：

* 它训练的视频数据集
* 要使用的基本度量和其他特性
* 回归器及其超参数

run\_training命令包含三个配置文件：一个数据集文件（包含有关训练数据集的信息）、一个特征参数文件和一个回归器模型参数文件（包含回归器超参数）。下面的示例代码定义了一个数据集、一组选定的功能、回归器及其超参数。

最后，可以扩展FeatureExtractor基类来开发定制的VMAF算法。这可以通过试验其他可用的基本度量和特性，或者发明新的度量和特性来实现。类似地，trainttestmodel基类可以扩展以测试其他回归模型。详情请参阅contribution.md。用户还可以使用现有的开源Python库，如SCIKIT学习（15）、CVXOPT〔16〕或TySoFrase[ 17 ]来使用替代的机器学习算法。软件包中包含scikit learn的随机林回归器的一个集成示例。

VDK软件包包括VMAF 0.3.1算法和基于NFLX视频数据集主观评分的训练支持向量机模型。我们还邀请社区使用该软件包来开发改进的特征和回归器，以便进行感知视频质量评估。我们鼓励用户在其他数据集上测试VMAF 0.3.1，并帮助改进我们的用例，并可能将其扩展到其他用例。

### 我们关于质量评估的公开问题

查看条件。Netflix支持数以千计的活动设备，包括智能电视、游戏机、机顶盒、电脑、平板电脑和智能手机，这导致我们的会员的观看条件大不相同。观看设置和显示会显著影响对质量的感知。例如，Netflix成员在4K 60英寸电视上观看以1 Mbps编码的720p电影时，如果在5英寸智能手机上观看，则可能对同一流的质量有非常不同的看法。目前的NFLX视频数据集涵盖了标准距离下的单一观看条件 - 电视观看。为了增强VMAF，我们在其他观看条件下进行主观测试。有了更多的数据，我们可以将算法进行推广，以便将查看条件（显示大小、与屏幕的距离等）输入到回归器中。

临时共用。我们当前的VMAF实现基于每帧计算质量分数。在许多用例中，希望在较长的时间内暂时将这些分数合并为一个值作为摘要返回。例如，一个场景的分数，一个常规时间段的分数，或者一整部电影的分数都是可取的。我们目前的方法是一个简单的时间池，取每帧值的算术平均值。然而，这种方法有“隐藏”劣质帧的风险。对较低分数给予更多权重的集合算法可能对人类感知更为准确。当使用摘要分数来比较帧之间不同质量波动的编码时，或者当优化编码或流会话时作为目标度量时，一个好的池机制尤其重要。对于VMAF和其他质量度量，一个感知上精确的时间池机制仍然是一个开放和具有挑战性的问题。

一致的标准。由于VMAF包含了完整的参考基本度量，VMAF高度依赖于参考的质量。不幸的是，Netflix目录中所有标题的视频源质量可能不一致。在我们的系统中，视频源的分辨率从SD到4K不等。即使在相同的分辨率下，最好的源可能会受到某些视频质量的影响。因此，比较（或总结）不同标题的VMAF分数可能不准确。例如，当从SD源生成的视频流达到VMAF分数99（满分100分）时，它绝不具有与从HD源编码的视频相同的感知质量（分数99）。对于质量监控，我们非常希望能够计算出各个来源之间一致的绝对质量分数。毕竟，当观众观看Netflix的节目时，除了交付到屏幕上的图片，他们没有任何参考。我们希望有一种自动化的方式来预测他们对交付给他们的视频的质量形成什么样的意见，同时考虑到导致最终呈现在屏幕上的视频的所有因素。

### 摘要

我们已经开发了VMAF 0.3.1和VDK 1.0.0软件包，以帮助我们在工作中向我们的成员提供最佳质量的视频流。我们的团队每天都使用它来评估视频编解码器、编码参数和策略，这是我们持续追求质量的一部分。VMAF和其他指标已经集成到我们的编码管道中，以改进我们的自动化QC。我们正处于使用VMAF作为客户端指标之一来监视系统范围的A/B测试的早期阶段。

在当今的互联网环境中，提高视频压缩标准和在实际编码系统中做出明智的决策是非常重要的。我们认为，使用传统的度量标准（并不总是与人类感知相关）会阻碍视频编码技术的真正进步。然而，总是依赖人工视觉测试是根本不可行的。VMAF是我们解决这个问题的尝试，使用我们内容中的示例来帮助设计和验证算法。与业界合作开发新视频标准的方式类似，我们邀请社区公开合作改进视频质量措施，最终目标是更有效地使用带宽，让所有人都能观看到令人愉悦的视频。

### 致谢

我们要感谢以下个人对VMAF项目的帮助：Joe Yuchieh Lin、Eddy Chi Hao Wu、C.-C Jay Kuo教授（南加州大学）、Patrick Le Callet教授（南特大学）和Todd Goodall。

### 工具书类

1. 王志刚，波维克，谢赫，西蒙切利，“图像质量评估：从错误可见性到结构相似性”，美国电气与电子工程师协会图像处理交易，第13卷，第4期，第600-6122004年4月。
2. BT.500：电视图像质量的主观评估方法，
3. M.-J.Chen和A.C.Bovik，“快速结构相似性指数算法”，《实时图像处理杂志》，第6卷，第4期，第281-287页，2011年12月。
4. N.Ponomarenko、F.Silvestri、K.Egiazarian、M.Carli、J.Astola和V.Lukin，“关于DCT基函数的系数间对比掩蔽”，《第三届消费电子视频处理和质量度量国际研讨会论文集》（VPQM'07），亚利桑那州斯科茨代尔，2007年1月。
5. 达拉编解码器。
6. T.-J.Liu、J.Y.Lin、W.Lin和C.-C.J.Kuo，“视觉质量评估：最新发展、编码应用和未来趋势”，APSIPA信号和信息处理交易，2013年。
7. 林俊杰，刘俊杰，吴永海，郭俊杰，“基于融合的视频质量评估（FVQA）指数”，APSIPA信号与信息处理交易，2014年。
8. C.Cortes和V.Vapnik，“支持向量网络”，机器学习，第20卷，第3期，第273-2971995页。
9. H.Sheikh和A.Bovik，“图像信息和视觉质量”，IEEE图像处理交易，第15卷，第2期，第430-4442006年2月。
10. S.Li，F.Zhang，L.Ma和K.Ngan，“通过单独评估细节损失和附加损害进行图像质量评估”，IEEE多媒体交易，第13卷，第5期，第935-9492011年10月。
11. S.Wolf和M.H.Pinson，“可变帧延迟（VQM VFD）的视频质量模型”，美国商务部，Nat。电信。信息管理，博尔德，公司，美国，技术备忘录TM-11-4822011年9月。
12. 视频质量专家组（VQEG），“高清视频内容视频质量模型验证报告”，2010年6月，
13. K.Seshadrinathan，R.Soundarajan，A.C.Bovik和L.K.Cormack，“视频主观和客观质量评估研究”，IEEE图像处理交易，第19卷，第6期，第1427-1441页，2010年6月。
14. A.K.Moorthy，L.K.Choi，A.C.Bovik和G.de Veciana，“移动设备上的视频质量评估：主观、行为和客观研究”，《信号处理选定主题的IEEE期刊》，第6卷，第6期，第652-6711912年10月。
15. scikit-learn: Machine Learning in Python. <http://scikit-learn.org/stable/>
16. CVXOPT: Python Software for Convex Optimization. <http://cvxopt.org/>
17. TensorFlow. <https://www.tensorflow.org/>

### See Also:

[**Per-Title Encode Optimization** delivering the same or better experience while using less bandwidthmedium.com](https://medium.com/@Netflix_Techblog/per-title-encode-optimization-7e99442b62a2)

[**High Quality Video Encoding at Scale** a video encoding pipeline that is highly robust, efficient and scalablemedium.com](https://medium.com/@Netflix_Techblog/high-quality-video-encoding-at-scale-d159db052746)

[**It’s All A/Bout Testing** The Netflix Experimentation Platformmedium.com](https://medium.com/@Netflix_Techblog/its-all-a-bout-testing-the-netflix-experimentation-platform-4e1ca458c15)

Originally published at [techblog.netflix.com](http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html) on June 6, 2016.