**目录**

**0. 摘要**

**1. 介绍**

**2. VMAF到HMAF**

2.1 VMAF

2.2 HMAF

**3. 主观测试思想和结果**

3.1创新的主观测试方法

3.1.1 传统的主观测试

3.1.2 创新的主观测试

3.2 实验场景的实现

3.2.1 实验设备介绍

3.2.2 人机交互原理

3.3 应用侧主观测试结果

3.3.1 单变量

3.3.2 双变量

3.3.3 四变量

3.4 网络侧主观测试结果

3.4.1 单变量

3.4.2 双变量

**4. HMAF结果对比和分析**

4.1 应用侧

4.1.1 数据处理

4.1.2 模型选择

4.1.3 结果

4.2 网络侧

**5. 总结**

**6. 参考文献**

**HMAF: Haptic Multi-indicator Assessment Fusion**

**摘要**

随着**5G**技术的发展，数据传输延迟持续降低，传输容量不断扩大。**扩展现实（eXtended Reality，XR）**等低延迟和高可靠性需求的业务已逐渐成为重要应用之一。XR业务在视听觉基础上提供触觉体验，促使研究者开始关注实现虚拟环境中玩家与玩家或环境之间多模态交互的应用场景。XR系统作为人类在环的闭环系统，**用户体验质量（Quality-of-Experience，QoE）**尤为关键，任何音、视、触觉方面的质量问题都可能影响用户的整体体验，因此需要系统性的评价指标来衡量交互质量。然而，现有的触觉评价指标（如HPW-PSNR、HSSIM、P-MSE）基于传输信号评估，尚不成熟且无法全面评估人类的主观感受，因此难以全面评价触觉的质量。

受到视频领域广受认可的评价指标Visual Multimethod Assessment Fusion（VMAF）的启发，我们创新性地将触觉的主观得分与客观参数相融合，**提出了Haptic Multi-indicator Assessment Fusion（HMAF）的方法思想**。此外，本文针对传统的主观测试方法在多种影响因素时评分效果较差的问题，**提出了一种更贴近真实结果的创新主观测试方法**。为了模拟测试场景，我们基于前沿的触觉XR手套设计了人机交互场景，并设计了**空域信息**（力反馈、振动、持续时间）和**时域信息**（端到端时延）的可调参数。研究结果显示，我们提出的HMAF方法在实际应用中具有重要意义。

**关键字**：HMAF，VMAF，多模态，XR，评价指标

**一、介绍**

触觉XR多模态应用形成了一个闭环系统，用户体验质量（QoE）显得尤为重要。无论是听觉、视觉还是触觉方面的质量下降，都直接影响到用户的整体感知。目前，音频领域的评价指标（比如PESQ和STOI）相对成熟，能够较好地评估语音清晰度等；而视频领域则采用了主观和客观相结合的多方法评估融合（VMAF），受到广大认可，它通过大量主观数据训练机器学习模型，将客观指标（如VIF、DLM和TI）与主观测试结果相结合，生成综合评估。然而，在触觉领域，尚缺乏类似的评价指标，急需一个可评估触觉交互质量的工具。

现有的传统触觉评价指标，如HPW-PSNR、HSSIM、P-MSE等，仍处于早期阶段。它们主要用于观察触觉数据在压缩前后的差异，未充分考虑网络传输中的时域问题。尽管这些指标考虑了主观因素，但它们是基于主观测试结果的客观测评，难以综合反映主观评价。因此，研究新的触觉评价指标，准确衡量用户的触觉感知质量，具有实际应用和研究价值。

受到VMAF将客观指标与主观评价融合的启发，本文设计了触觉XR人机交互应用场景。通过主观测试收集数据，考虑了应用侧的力反馈、振动强度、振动持续时间以及网络传输的端到端延迟。最终，利用机器学习模型训练了触觉多指标评估融合（HMAF）评价指标。这种新型指标优于目前主流指标，更贴近用户主观评价。

此外，本文还提出了一种创新的主观测试方法。传统的方法将多个影响因素整合为一个模式，测试人员需要随机选择一个模式进行评分，但难以在考虑大量影响因素的情况下收集准确的主观数据。因此，本文提出一种让测试人员在理想情况下（如无端到端延迟）自由调整多个影响因素值来评分的创新方法。经验证，这种方法更贴近真实场景。

最后，本文还应用HMAF设计了一个网络侧模型，即在网络传输中调整数据包丢失率（PLR）、延迟和Jitter。网络侧模型结果同样验证了HMAF思想的正确性和可行性。

该论文共分为五个部分：第一部分介绍；第二部分受VMAF启发的新触觉评价指标HMAF；第三部分介绍主观测试方法；第四部分充分讨论HMAF的结果；最后，对未来工作进行总结和规划。

**二、VMAF到HMAF**

**2.1 Visual Multimethod Assessment Fusion**

Netflix提出了一种名为VMAF的视频评价指标，它结合了主观和客观的评价。在评估视频质量时，VMAF考虑了视频的时间和空间特征。具体来说，VIF和DLM用于从视频帧中提取空间域特征，TI用于提取捕获多帧间相关性的时域特征。通过对提取的时间和空间特征进行处理，可以综合评价潜在特征和失真类型。

如图1所示，首先，VMAF进行像素级空间特征提取（VIF和DLM）和时间特征提取（TI），可以反映视频帧的质量。然后，对提取出的特征进行帧级的帧内空间池化处理。这两步模拟计算机对人体视觉系统的神经处理，生成视频评价的客观指标。

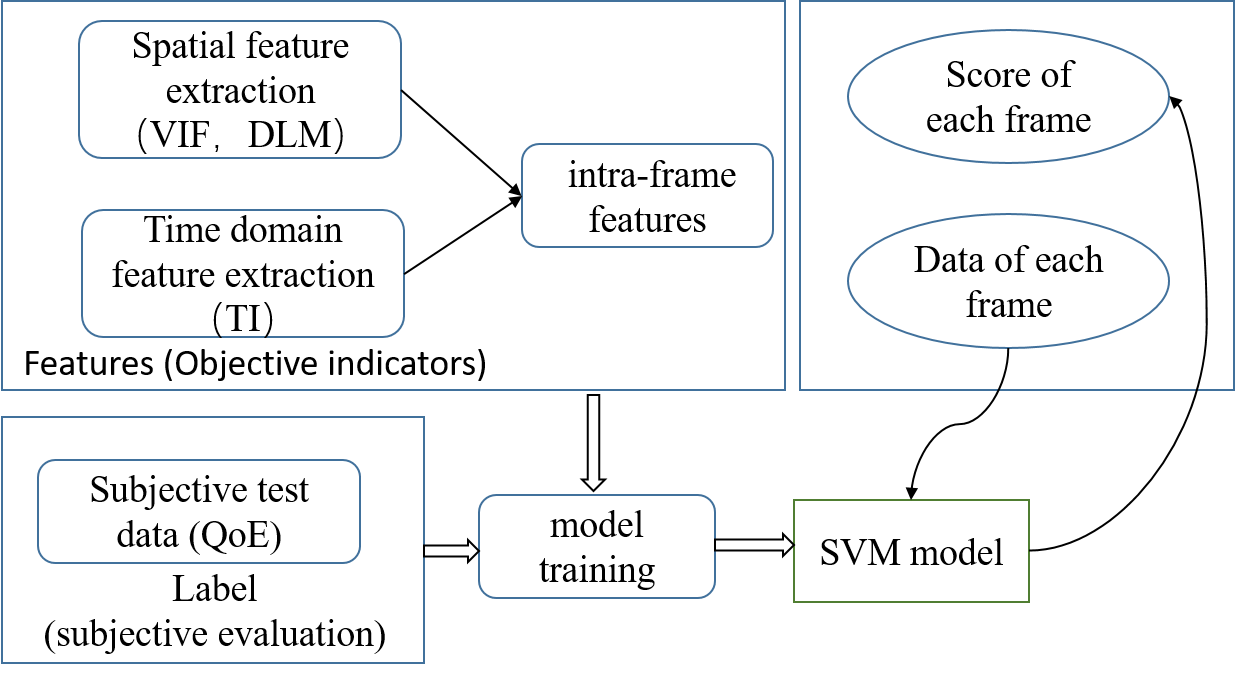


图1. VMAF

VMAF通过让测试对象参与主观实验来收集主观评价数据，以收集他们对视频帧的主观质量评价。用户提供了一个5分的视频帧的质量，从“差”到“优秀”不等。然后，测试对象在同一帧下的平均得分，以获得平均意见得分（MOS），并使用5分参考量表映射到VMAF量表上。一般来说，“差”在VMAF量表上是20分，而“优秀”在VMAF量表上是100分。然后将映射出来的数据用于训练，VMAF的预测分数将落在0到100分之间。

最后，VMAF采用机器学习方法进行分类预测，将不同权重的指标结果融合，得到客观指标和主观实验融合的评价得分。利用监督学习方法对机器学习模型进行训练和测试，并生成一个评价模型。

**2.2 Haptic Multi-indicator Assessment Fusion**

受到VMAF的启发，本研究也着重考虑了空间和时间特征。我们采用了两种手套：触觉手套“Dexmo”和振动纹理感应手套“AvatarVR”。空间特征包括AvatarVR引发的每个手指振动的强度和持续时间，以及Dexmo提供的力反馈。而时域特征则是应用层传输过程中的端到端时延。

在实验中，我们模拟了不同场景下振动纹理手套的不同振动强度和持续时间，以及触觉手套的不同力反馈情况。假设通信网络质量良好，没有丢包或其他问题，只存在通信延迟。我们通过呈现不同场景给用户进行主观测试，为用户提供了从“差”到“优秀”的5分评分量表。

最终，我们采用机器学习算法对数据进行拟合，得出了融合了主观评分和客观指标的综合评价指标。这些客观指标不仅包括了单独的空间或时域特征，还考虑了两者的综合效果。图2展示了HMAF的实现框架。

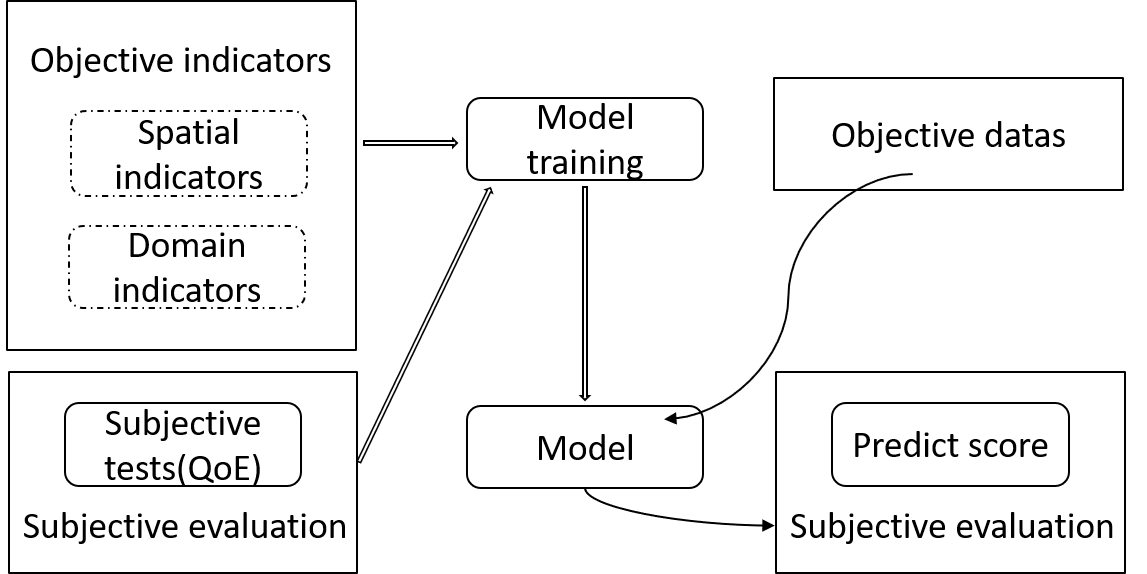


图2. HMAF

**三、主观测试**

本实验分别采用传统的主观测试方法和提出的创新的主观测试方法，分别对应用侧和网络侧进行了测试。用户可以在应用侧调整的影响因素分为空间域和时域。空间域包括力反馈、振动强度和振动持续时间。其中时域包括传输期间的端到端延迟。网络侧是专门设置以探讨网络参数对用户体验的影响。所有可调节的影响因素包括数据包丢失率（PLR）、网络端到端延迟（Delay）和用来模拟真实网络波动的Jitter设置。主观测试分别为单变量测试、双变量测试和四变量测试。其中，单变量测试探讨单一影响因子与QoE之间的关系，而双变量和四变量用于训练HMAF生成评价模型。在进行测试之前，使用3D打印作为力反馈Ground Truth，制作了一个真实的3D球。此外，还设计了数据采集传感器，建立了传感器数据与用户界面参数之间的映射关系。在每个测试者调整UI界面之前，他们首先抓住真正的3D球来收集每个用户GroundTruth。如下图3所示

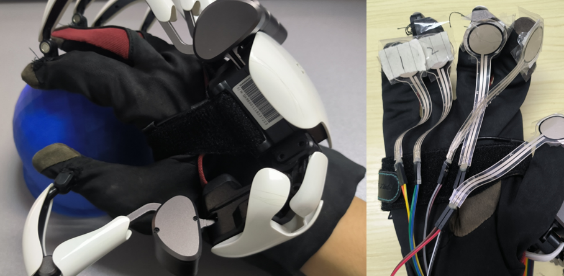


图3. GroundTruth-3D球和数据采集传感器

主观测试为用户提供了一个五级评分参考。分别有“bad”、“poor”、“fair”、“good”和“excellent”。“excellent”对应5分，对应最接近真实3D打印球的触摸感觉；“good”对应4分，对触觉体验有轻微影响，但不影响正常体验，仍能正常互动；“fair”3分，对触觉互动有显著影响，已经影响了测试者的正常判断，即使仔细区分也很难体验真实的触摸；“poor”2分，即几乎失去触觉体验；“bad”1分，对应完全扭曲或无法感觉到力反馈。

**3.1 创新的主观测试方法**

本实验为主观测试设计了两种不同的UI界面，如图4和图5所示。图4显示了传统的主观测试方法的界面，图5显示了创新的主观测试方法的界面。

***3.1.1 传统的主观测试方法***

传统的主观测试中，测试前设计多种模式，进行不同数值的影响因子的组合。受试者随机选择多种模式中的一种，并从1到5分进行评分。然后再随机选择其他组合模式，并逐个给他们打分。如图3所示，本实验也使用了传统的主观测试方法。然而，传统的主观测试方法的结果是随机的和混乱的，特别是当触觉反馈存在多个影响因子时，与真实结果的偏差很大。收集到的测试结果也证实了传统的主观测试方法的缺点。

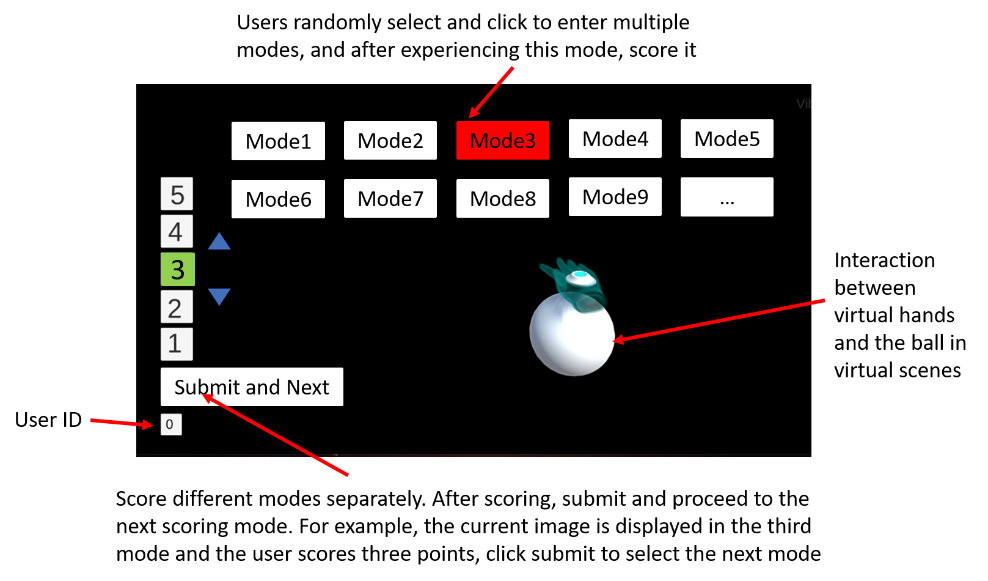


图4. 传统主观测试方法

***3.1.2 创新的主观测试方法***

为了解决传统主观测试结果中存在较大偏差和随机混乱的问题，本文提出了一种创新的主观测试方法，如图5所示。告知受试者当前最优场景，并让他们分别自由地调节每个影响因素的slider。通过自由滑动每个影响因子slider，用户可以找到他们认为与每个分数对应的5-1分，并保存所有的影响因子值。通过数据收集比较，我们创新的主观测试方法比传统的主观测试方法更接近用户的真实体验。这种主观评价数据在应用于机器学习训练分类时将更加准确。

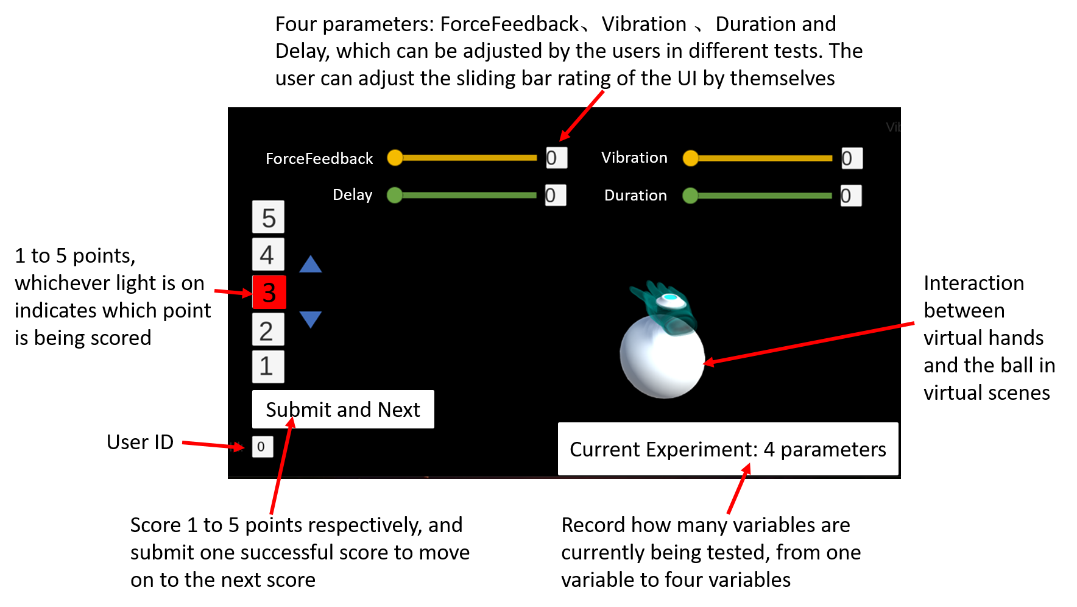


图5. 创新主观测试方法

**3.2 实验场景的实现**

***3.2.1 实验设备***

研究对象为虚拟现实触觉XR多模态感知交互，使用了两种XR手套。图6显示了用于主观测试的设备，其中包括动觉手套Dexmo和振动纹理觉手套AvatarVR。Dexmo手套使用类似外骨骼的结构，通过设置不同的力反馈，获得相应的力反馈，模拟与不同硬度物体的相互作用和模拟触觉反馈。AvatarVR手套是一种振动纹理觉手套，在五个手指和手掌上都有振动点，可以设置不同的振动强度和振动持续时间（振动一次的时间），通过振动模拟纹理感知。



图6. 实验设备

然而，仅仅力反馈或仅仅振动反馈为受试者提供了的主观感受相对单一简单，受试者不能充分体验与虚拟场景的互动。因此，本实验创新性地将两种手套结合在一起，让受试者在互动过程中同时体验到触觉反馈和振动纹理感知，丰富他们的体验。此外，使用HTC Vive HMD头盔与虚拟场景进行视觉交互，使用跟踪器Tracker定位3D空间信息。

***3.2.2 人机XR/VR交互***

如图7所示，本实验设计了一个人机虚拟现实交互场景，以虚拟场景中的一个球作为交互对象。为了确保网络传输延迟在控制下，建立了客户端和服务端，并使用网络模拟器来控制网络延迟。客户端与手套、头盔等触觉设备连接，通过Tracker获取手套的3D空间位姿信息，将其发送到服务端。在获取位姿信息后，服务器会更新渲染场景中受控模型手的位姿，并与场景中的小球进行交互。交互后，模型手与球之间产生反馈信号，通过下行传输发送反馈信号和视频信息回客户端，客户端根据接收到的信息控制真实XR手套产生力反馈和振动，用户通过VR头盔观看传输视频。

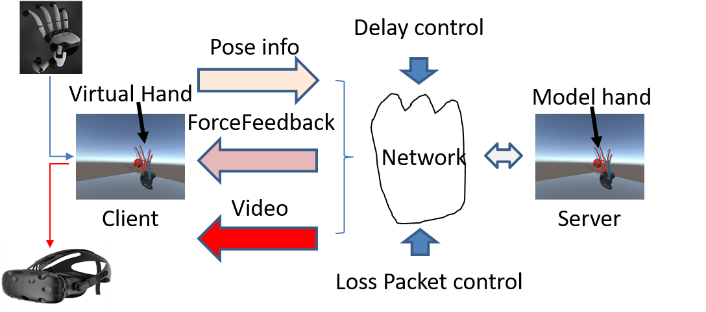


图7. 人机虚拟现实交互示意图

**3.3 应用侧主观测试设计和结果**

***3.3.1 单变量测试结果***

如图5所示为应用测主观测试UI界面。应用侧的单变量测试分别探讨了QoE与力反馈、振动强度、振动持续时间和端到端传输延迟之间的关系。力反馈和延迟分别使用Dexmo进行了测试。使用AvatarVR测试振动强度和持续时间。在进行单变量测试时，本实验坚持控制变量的原则。例如，测试力反馈时，保持端到端传输延迟在0 ms；测试延迟时，控制力反馈为用户单独测试值5分的均值不变。

此外，在测试延迟时，本实验测试了端到端时延的Just Noticeable Difference (JND)，即将受试者刚刚感到时延影响的阈值。该JND对网络性能要求具有重要的参考意义。

本实验共对30人进行了单变量测试，得到了QoE与力反馈、振动、持续时间和延迟之间关系的图表，如图8a-d所示。红点表示该评分下所有受试者调节影响因子数值的平均值，绿点表示所有样本值。通过数据分析和问卷调查，得出的基本结论如下：

* 力反馈：力反馈间隔映射设置为0-100（2N）。受试者能够感受到力反馈的最低阈值在5到15分（主观评价1分）之间。他们认为最接近接触真实3D小球触感的5分值在50到70之间。此外，QoE和力反馈基本上是线性相关的。
* 端到端时延：端到端时延的JND为30-60 ms。受试者一般认为在160 ms以上的延迟是不可接受的。
* 振动强度和振动持续时间：设置振动强度范围为0-1。振动强度和持续时间越大，用户认为的体验质量就越差，受试者认为感受会逐渐失真。同样，这两个影响因子与QoE也基本呈线性相关。

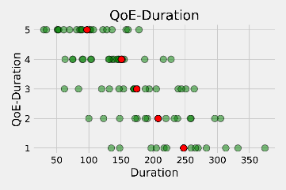
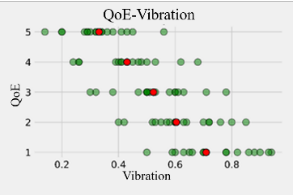
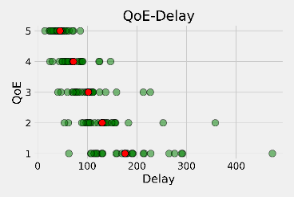
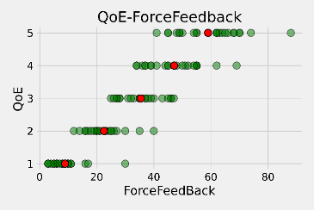


图8a-d. 单变量测试中QoE-单个影响因子关系

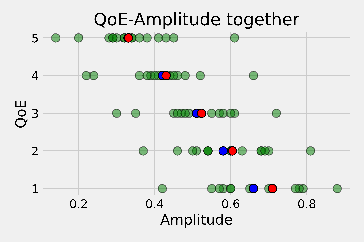
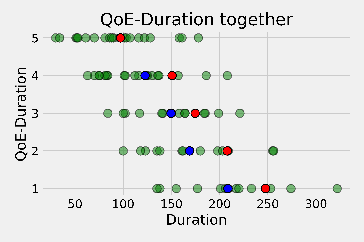
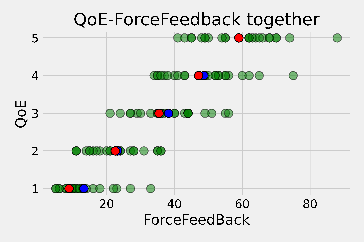
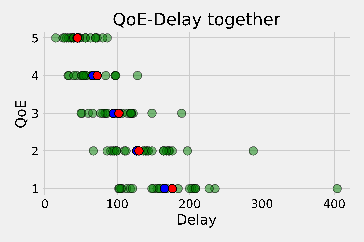


图8e-h 双变量测试中QoE-单个影响因子关系

***3.3.2 双变量测试结果***

实验采用动觉手套Dexmo和振动纹理觉AvatarVR手套的融合手套进行双变量测试。测试组合分别为：力反馈和延迟、振动强度和持续时间。将双变量分为以上两组测试的原因如下：振动纹理感知手套AvatarVR设置了不同的振动和持续时间来模拟纹理感知。当涉及到延迟时，特别是当延迟小于持续时间时，延迟的感觉会被持续的振动所覆盖，因此AvatarVR无法测试延迟。相反，在Dexmo手套中的每个力反馈的设置不需要是连续的，Dexmo可以用来测试延迟。双变量主观测试同样也采用了控制变量的原理。本部分实验收集了30组双变量主观测试数据。

本部分得到了QoE与延迟和力反馈、QoE与振动和持续时间之间的关系。双变量测试中QoE-各单个影响因子之间的关系如图8e-h所示。红点表示单变量测试中该影响因子的平均值，蓝点表示双变量中该影响因子的平均值，绿点表示该影响因子在双变量测试中的样本点。通过数据分析和问卷调查得出的基本结论如下：

* 两个影响因子同时影响用户体验质量时，当共同调节这两个影响因子时，这两个因子的变化将同时加剧QoE的变化。如图8e所示，双变量测试中延迟的曲线（蓝色）低于单变量测试中的延迟曲线（红色）。这说明，当两个影响因子共同变差时，QoE的质量将会加剧变差。其余影响因子同理。
* 通过主观调查问卷：在力反馈与延迟相结合的情况下，多数用户（27/30）认为力反馈对体验质量的影响更大，即力反馈的权重更高；在振动和持续时间结合时，多数用户（26/30）认为振动强度的影响更大

此外说明，以上双变量测试和下述四变量测试都用于HMAF模型的训练。

***3.3.3 四变量测试结果***

本部分收集了17组实验数据进行四变量测试。四变量测试是指用户同时调整力反馈、振动、持续时间和延迟四个变量。由于单个变量与QoE之间的关系已经在之前的实验中得到了探索，因此收集这组实验数据的目的是训练机器学习模型生成多个数据，形成最终的评价指标模型。

**3.4 网络侧主观测试设计和结果**

网络侧是专门设置用以探讨网络传输参数对用户体验的影响，并用以验证HMAF思想模型的普适性。网络侧UI界面如图9所示。为了控制网络，将其设置为通过网络模拟器控制客户端和服务器之间的端到端时延和丢包。在测试中分别调节三个影响因素： Packet Loss Rate（PLR）、端到端时延和Jitter。网络侧测试依赖于上一节应用侧测试的结果。从应用侧测试中获得的JND、力反馈、振动和持续时间，在网络侧场景中被分配并保持固定。

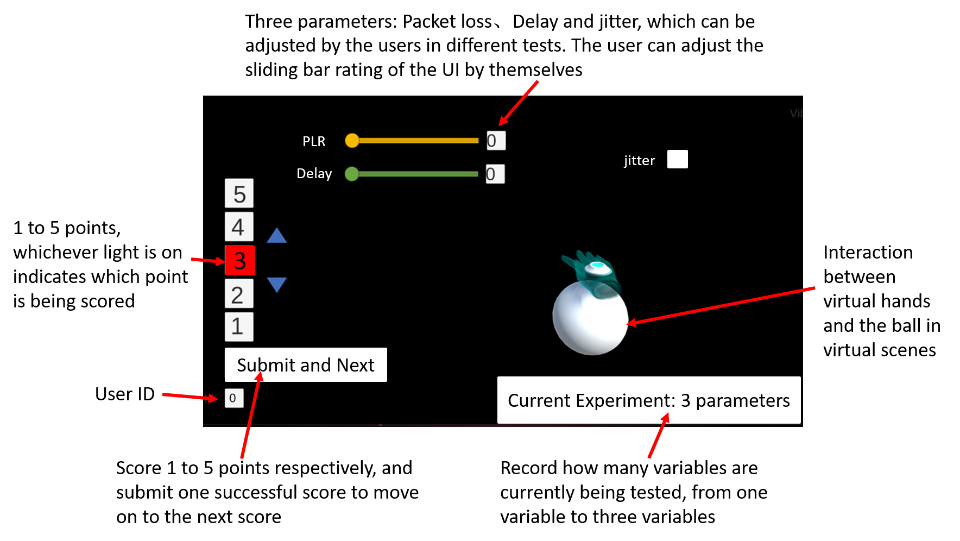


图9. 网络侧主观测试UI界面

***3.4.1 单变量测试结果***

本实验分别收集了31组单变量数据。网络侧单变量实验测试了QoE与端到端延迟、QoE与PLR以及QoE与Jitter之间的关系。与应用侧测试类似，实验保持了控制变量的原则。例如，在测试端到端时延时，Jitter不打开，也保持PLR等于0；在测试PLR时，没有端到端延迟。其中数据丢包相关设置如下：触觉信号无编码，丢包后应用层无补偿，丢包方式为随机丢包模式。

通过单变量主观测试，得到了QoE与PLR之间的关系、QoE与端到端时延之间的关系、QoE与Jitter之间的关系，如图10a-d所示。

图10a-c中红点表示单变量测试中该影响因子在1到5分的平均值，绿点表示该影响因素的样本值。图10d是打开Jitter前后的延迟的比较。红点表示Jitter打开后的平均得分，蓝点表示Jitter未打开时的平均得分。通过数据分析和问卷调查得出的基本结论如下：

* 网络侧的端到端时延与QoE之间的关系与应用侧的QoE-端到端时延关系基本一致。
* 通过调查问卷：在Jitter被打开后，延迟越小，Jitter的影响就越大。此外，在时延和PLR中，大多数人（20/31）认为时延的影响更大，而少数人（11/31）认为丢包的影响更大。

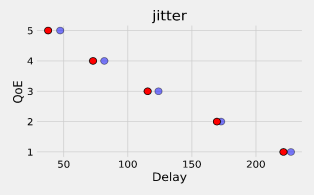
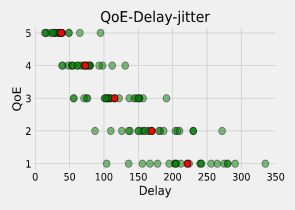
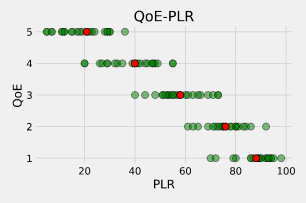
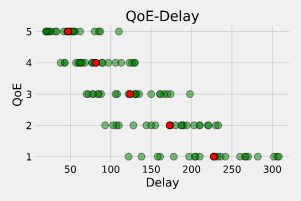


图10a-d. 网络侧单变量测试结果

***3.4.2 双变量测试结果***

双变量测试分别收集了30组主观测试数据。双变量测试对QoE与Delay和PLR之间的关系进行了测试，以及QoE与Jitter和PLR之间的关系。随着Jitter打开或关闭，受试者同时调整延迟和PLR的slider进行主观打分。在单变量检验中已经探索了网络侧单变量与QoE之间的关系，因此收集这组实验数据的目的是训练多数据的机器学习模型，在网络侧生成最终的评价指标模型。验证提出的HMAF思想也可以在网络侧环境中得到验证。

**四、HMAF结果**

**4.1 应用侧**

***4.1.1 数据处理***

应用侧采集了双变量和四变量的主观数据。双变量测试包括：QoE与力反馈和延迟之间，QoE与振动和持续时间之间的测试。四变量测试为QoE和力反馈、延迟、振动、持续时间之间的测试。使用箱形图去除每一组数据中的异常值。

主观测试评分从1分到5分，1分代表最差，5分代表最好。为与传统的触觉评价指标P-MSE模型进行映射，本文将得分归一化为0-100分，即1对应0,2对应25,3对应50,4对应75,5对应100。

***4.1.2 模型选择***

本文选择机器学习的9种回归算法，包括决策树回归、线性回归、SVM回归、KNN回归、随机森林回归、Adaboost回归、GBRT回归、Bagging回归和 ExtraTree回归。通过实验比较，我们选择了拟合效果最好的GBRT回归算法作为模型。具体选择模型的方法：将数据集分为两部分，80%作为训练集，20%作为测试集。每个模型训练1000次，计算1000个预测结果的平均R2值，选择值最高的模型。

对于P-MSE，计算公式1所示，其中为4，表示i-th变量原始值，表示i-th变量的实际值，c表示每个变量的权重。由于P-MSE的最优值为0，为了便于比较，我们对其进行了映射。其预测值为100-P-MSE。我们通过梯度下降算法得到了c。首先，我们通过遍历得到RMSE较小的点，然后将该点作为梯度下降的起点。最后的表达式是：

其中，a为16.5721，b为14.6151，c为23.7967，d为0.2175，D为延迟，F为力反馈，Dur为持续时间，V为振动。当预测值小于0时，将预测值设置为0。

***4.1.3 应用侧双变量结果***

对于双变量主观测试数据，分别使用HMAF和P-MSE进行预测，分别计算RMSE和PLCC，计算结果如表1所示。

|  |  |  |
| --- | --- | --- |
|  | Proposed model | P-MSE |
|  | 0.9769 | -0.3200 |
| RMSE | 6.6705 | 42.8718 |
| PLCC | 0.9895 | 0.5709 |

在验证了性能指标后，HMAF的PLCC大于0.95，RMSE小于10。与P-MSE相比，提出的指标具有可信度和优越性。通过对预测的分数进行反向映射，即将分数重新映射到1-5分，绘制了以下混淆矩阵和分类图。从图11中可以看出，HMAF远远优于PMSE模型。

图12显示了双变量模型的对应预测错误标记。左侧为HMAF模型，右侧为P-MSE模型。蓝色表示正确的分类，橙色表示错误的分类。可以看出，HMAF预测结果没有分类误差，远远优于P-MSE模型。

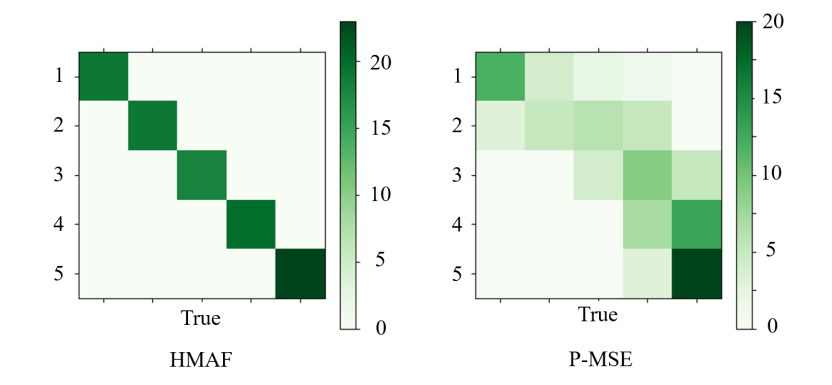


图11. 应用测双变量预测结果混淆矩阵对比

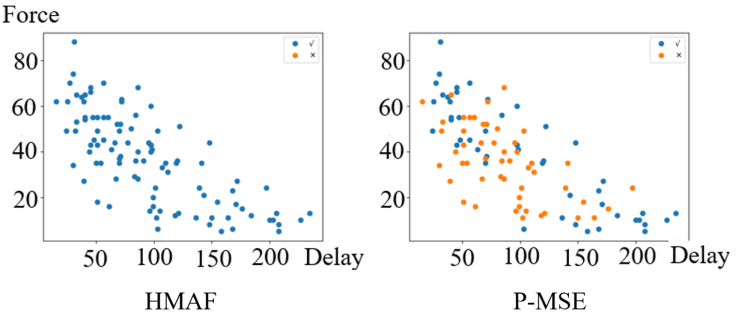


图12. 应用测双变量预测结果错误分类图对比

***4.1.4 应用侧四变量结果***

对于四变量主观测试数据，分别使用HMAF和P-MSE进行预测，分别计算RMSE和PLCC，计算结果如表2所示。同样，HMAF远优于P-MSE。同样，通过将预测的分数进行反向映射，即将分数重新映射到1-5分，绘制了以下混淆矩阵和分类图。从图13中可以看出，HMAF四变量模型也远远优于PMSE模型。

|  |  |  |
| --- | --- | --- |
|  | Proposed model | P-MSE |
|  | 0.9436 | 0.2112 |
| RMSE | 8.6801 | 35.6491 |
| PLCC | 0.9757 | 0.7352 |

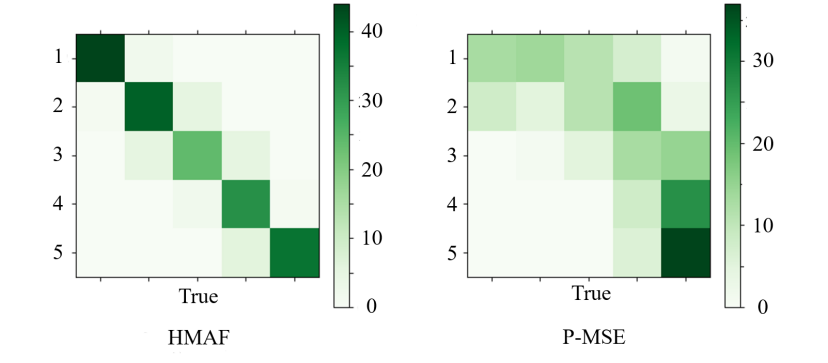


图13. 应用测四变量预测结果混淆矩阵对比

**4.2 网络侧**

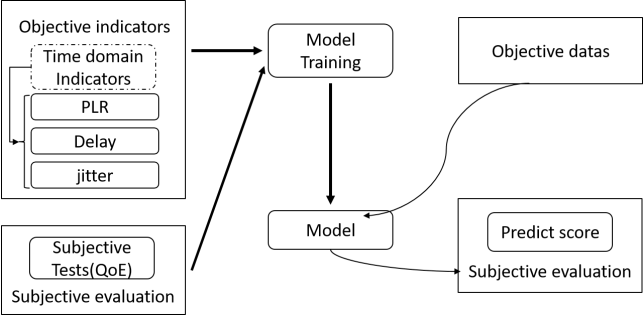


图14. 网络侧HMAF原理图

上图14为变化为网络侧的HMAF思想原理图。网络侧专门设计用于探索网络性能对触觉交互的影响，同时也用于HMAF思想模型的时域特性的验证，网络侧主观测试专门收集了网络性能相关的主观测试数据，包括两个双变量模型：QoE与Delay和PLR之间，QoE与PLR和Jitter之间的关系。

为了后续的结果显示，并与传统的触觉评价指标P-MSE模型构建进行比较，网络侧也将1-5分归一化为0-100分。对于异常值，仍使用箱形图来清除。网络侧同样使用机器学习中9种回归算法。通过实验比较，选择了拟合效果最好的决策树回归算法作为模型。具体的选择方法与应用侧的方法相同。

网络侧HMAF和P-MSE预测结果如下图15和图16，分别计算RMSE和PLCC，计算结果如表3所示。同样逆映射得到的混淆矩阵分类图，如图15和16所示。

|  |  |  |
| --- | --- | --- |
|  | Proposed model | P-MSE |
|  | 0.9919 | 0.6264 |
| RMSE | 6.3066 | 21.4702 |
| PLCC | 0.9961 | 0.9077 |

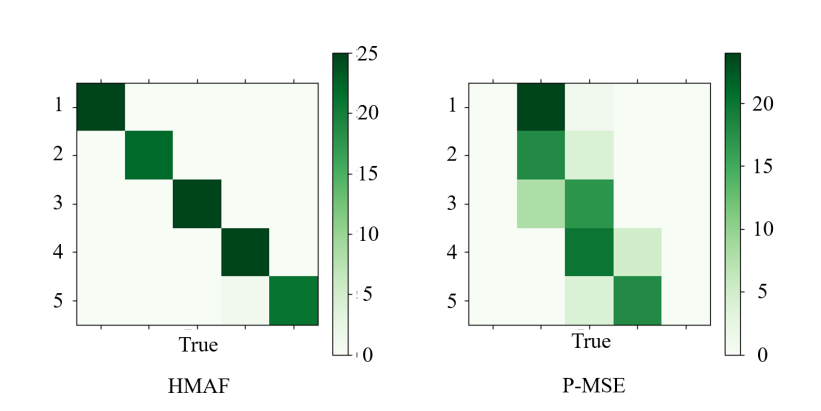


图15. 网络侧双变量预测结果混淆矩阵对比

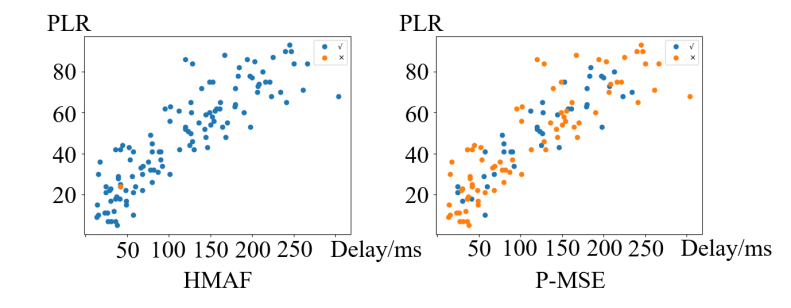


图16. 网络侧双变量预测结果错误分类图

通过结果比较，可以发现在网络侧使用HMAF思想也可以很好地对双变量模型进行预测，与传统的P-MSE相比有了显著的改进。验证了在仅具有时域特征的网络侧应用HMAF思想是完全可行的，验证了该思想的正确性。

**五、总结和未来工作**

本文创新性地提出了一种利用机器学习方法将主观触觉评价与客观参数相结合的训练思想。采用创新的主观测试方法，获得更接近现实的主观数据。本文设计并实现了人机虚拟现实交互的场景，并通过调整时空特征影响因子，探索了各种影响因子与QoE之间的关系。通过将HMAF与传统的触觉指标进行比较，我发现HMAF优于现有的传统评价指标，更接近受试者的主观评价，这也为后续的触觉评价指标的开发提供了创新的思路。同样，本文也在网络侧进行了研究，我们的模型思想在网络侧也同样得到了验证，这表明了HMAF思想的普适性。未来的工作将把应用侧和网络侧结合起来，将更多的时空信息集成到一个模型中。