媒体主观体验质量的分布规律

朱文瀚[1],张晓菁[1],翟广涛[1*],张晓平[2*]
（[1]上海交通大学，图像通信与网络工程研究所，上海 200240
[2]瑞尔森大学，电子与计算机工程系，加拿大）

摘要：随着移动多媒体技术的迅速发展，数字媒体内容开始充斥并丰富着我们的日常生活。人们可以通过手机终端及其他智能设备便捷即时地获取和分享数字图像。然而在数字媒体的传输和处理过程会不可避免地导致原始的高分辨率图像变得模糊进而出现失真，因此使用媒体体验质量评价来监督和评估数字媒体中内容处理的效果十分重要。由于人眼是视觉信号的最终获取者和接受者，因此精确有效的主观媒体体验质量评价结果能够为测试、训练和评估客观质量评价算法提供可靠的标准。目前已有许多图像质量评价数据库提供了关于失真图像的评分用以指导客观评价算法，这些数据库均使用平均意见得分或平均意见得分差异作为评价指标。但是仅用一个平均数值来反映媒体内容的主观整体质量有失合理性和准确性，可以考虑得更为全面，例如媒体主观体验质量分数服从某种分布，从而对现有的主观质量评价的评价指标进行改进和完善。本文基于 LIVE 数据库中的部分图像重建了新的媒体主观体验质量数据库，研究了媒体主观体验质量分数及其统计量与评分分布类型、图像失真类型和图像内容的关系，为媒体主观体验质量的进一步研究提供了参考依据。

关键词：媒体主观体验；主观质量评价；数据分布；假设检验；平均意见得分
中图分类号：O422　　**文献标识码：A**

Distribution Regularities of Media Subjective Experience Quality

Wenhan Zhu[1], Xiaojing Zhang[1], Guangtao Zhai[1*],Xiaoping Zhang[2*]

([1]Shanghai Jiao Tong University, Institute of Image Communication and Network Engineer,Shanghai,200240,China

[2]Ryerson University, Department of Electrical and Computer Engineering, Toronto, Canada)

Abstract: With the rapid development of mobile multimedia technology, digital media contents have become ubiquitous and are flooding and enriching our daily lives. Digital images can be easily and instantly obtained and shared through mobile phones and other smart devices. However, it is inevitable that procedures such as image acquisition, compression, transmission and restoration in digital media will contribute to distortion of original images with high resolutions, thus rendering media experience quality assessment to monitor and evaluate the effect of processing in media particularly important. At the same time, since humans are the ultimate receivers of visual signals, effective and accurate subjective experience quality of media is able to provide reliable standards for the testing, training and evaluating of objective quality assessment algorithms. At present, there are a number of image quality assessment (IQA) databases that provide subjective scores on distorted images, which use mean opinion score (MOS) and differential mean opinion score (DMOS) as evaluation indicators. Nevertheless, only using an average score to reflect the subjective overall quality of media content is lack of rationality and accuracy, which should be considered more comprehensively. For instance, the media subjective

experience ratings are subject to a certain distribution, so it is necessary to improve the existing subjective quality evaluation methods. In this paper, we construct a new media experience quality assessment database based on materials of the LIVE database. Also, we investigate the relationship between statistics of subjective ratings and the type of score distribution, categories of image distortion and image content. Furthermore, this work provides a reference for further research on the subjective experience quality of media.

Key words： Media subjective experience; subjective quality assessment; data distribution; hypothesis testing; mean opinion score

1 引言[1]

随着网络通信与信号处理技术的不断发展，数字媒体中的图像和视频的数量大幅上升，在游戏娱乐、测控遥感、医学诊疗、交通管理和安全服务等多方面都得到了广泛的应用和深入的研究[1]。由于数字媒体在获取存储内容、传输适配等实际应用的过程中会不可避免地造成媒体内容的失真和降质，从而产生了保持或提升媒体体验质量的需求，来降低这些视觉可感知的媒体体验质量损失[2]、使得人们能够有更好的视觉体验。因此，我们需要通过媒体体验质量评价来量化媒体内容的退化程度，以便于对媒体体验质量进行进一步的优化处理，也能更好地将图像处理技术应用于数字媒体传输、成像以及识别和安全监控等领域，促进各技术协同发展[3]。

图像质量评价是通过对图像质量进行定量分析来评估图像的失真情况和优劣程度，分为主观评价方法和客观评价方法[4]。其中，客观评价方法是一种基于计算机的方法，通过建立某种数学模型来模拟人类视觉系统（Human Visual System, HVS）从而根据计算公式对图像的整体质量进行预测和自动评价；而主观评价方法是一种基于人的方法，依据自己对图像整体视觉质量的主观感受进行评价，评价过程中需要一组测试人员结合评分标准和规则对一系列图像的质量进行判断和打分，最后对所有测试者给出的意见进行处理并得到最终评分[5]。

由于在大多数多媒体应用中，人类观察者是视觉信息的直接感知者和观测者，也是最终使用者和解释者[6]，因此评估媒体体验质量的标准是人的主观判断，是验证优化模型的唯一方式，也是最为直接与可靠的媒体体验质量评价方法。

主观的媒体体验评价方法一般采用的评估指标为 MOS 和 DMOS。MOS 值越大，DMOS 值越小，说明目标内容质量越高。目前，许多主流的主观媒体体验质量评价研究主要是通过建立图像数据库进行的，以下 5 个开源数据库是目前使用最为广泛且使用频率最高的图像质量评价数据库：LIVE[7]、CSIQ[8]、IVC[9]、TID2008[10]、TID2013[11]。前两个数据库提供的数值形式为 DMOS，后三个为 MOS，都是以一个平均数值来代表具体某张图像的主观评分值[12]；例如，LIVE 数据库[7]应用最为广泛，包含 29 张参考图像和 779 张失真图像，所有图像皆为彩色图像。有 5 种失真类型，包括 JPEG 压缩、JPEG2000 压缩、白噪声、高斯模糊和快速衰落。每种失真类型包含 4 或 5 种失真等级。图像分辨率有多种，例如 480×720、610×488、627×482、634×438、768×512。该数据库的评价指标为 DMOS 值，由 161 个观察者打分得到，DMOS 取值范围为[0,100]。TID2013 数据库[11]是 TID2008 的加强版，包含 25 张参考图像和 3000 张失真图像，所有图像皆为彩色图像。有 24 种失真类型，增加了如有损压缩、彩色图像量化等 7 种失真类型。每种失真类型包含 5 种失真等级。图像分辨率为 512×348。该数据库的评价指标为 MOS 值，由 971 个观察者打分得到，MOS 取值范围为[0,9]。CSIQ 数据库[8]包含 30 张参考图像和 866 张失真图像，所有图像皆为彩色图像。有 6 种失真类型，包括 JPEG 压缩、JPEG2000 压缩、高斯模糊、加性高斯白噪声、加性高斯粉红噪声和整体对比度缩减。每种失真类型包含 4 或 5 种失真等级。图像分辨率为 512×512。该数据库的评价指标为 DMOS 值，由 25 个观察者打分得到，DMOS 取值范围为[0,1]。

基金项目：中国国家自然科学基金 NSFC61831015
作者简介：朱文瀚（1993-），男（汉族），江苏南京人，上海交通大学大学博士研究生，zhuwenhan823@sjtu.edu.cn。
通讯作者：翟广涛（1978-），男（汉族），山东济南人，上海交通大学教授，zhaiguangtao@sjtu.edu.cn。
通讯作者：张晓平，男，瑞尔森大学教授，xzhang@ee.ryerson.ca。

然而，在媒体体验质量主观评价的描述分数上有进一步优化的空间。研究表明 MOS 具有主体同质性的隐含假设[13]，因此仅使用 MOS 值分析主观测试结果会遗漏用户评分的多样性信息以及数据背后一些潜在的重要特征[14]，例如测试过程中常见的宽大效应和近因效应，同时，一个实体的主观平均评分的分布会由于选择性偏差这一用户行为特点而与潜在质量分布有所区别[15]。此类现象也可运用于媒体体验质量主观评价，所有用户或观察者对每一张图像的评分应当服从某一种分布。由于 MOS 只是有限数量个人评分分布的统计平均数，因此只是反映该评分分布的众多统计量中的一个，所以仅用一个数值来反映图像质量的主观评分缺乏合理性。因而媒体体验质量评分的进一步合理化以及对于媒体体验主观质量分数统计特性的研究可以改善媒体体验质量数据库的准确性，也能为客观评价提供更可靠的参考依据。

依据测试过程中是否给出原始图像作为参考标准，主观评价方法可分为两类：绝对评价和相对评价[16]。绝对评价是在无参考图像的情况下根据已知的评价准则或个人经验直接对图像质量进行视觉感受上的评分，评价指标主要是 MOS，通过对图像质量由好到坏进行尺度划分并与评分数值相对应，该李克特量表称作“全优度尺度”[17]，通常来说其中的妨碍尺度和质量尺度分别给专业人士和非专业人士评分使用。相对评价是将一批图像进行质量的相互比较、优劣排序及评分，评价指标主要是 DMOS，该 5 分制评分称作“群优度尺度”。实际应用中，除了使用对应等级的离散的评分数值，也可使用一定范围内。

国际上提出了多种媒体体验质量主观评价方法的标准，主要分为四类：单激励法、双激励法、强迫选择量表法（迫选法）和相似度对比法（成对比较法）[18,19]。其中，单激励法和双激励法的区别在于需要判断和评分的图像是单个还是一对[20]。单激励法中常用的是单激励连续质量量表法（Single Stimulus Continuous Quality Evaluation, SSCQE）。双激励法中常用的是双激励连续质量量表法（Double Stimulus Continuous Quality Scale, DSCQS）和双激励损伤量表法（Double Stimulus Impairment Scale, DSIS）。双激励损伤量表法主要用于测量系统的降质特性[21]。在该种交替的方法中，给定原始图像（未失真的参考图像）和待测图像（失真的测试图像）组成的“图像对”，先显示原始图像，后显示失真图像，观察者通过观察和对比这样的一系列图像对后，根据图像主观质量 5 级损伤量表给待测图像评分。该方法更适用于整个的损伤范围。双激励连续质量量表法主要用于测量系统相对于某一基准的质量[21]。所用方法同样是交替方法，但与双激励损伤量表法有所不同，在该种方法中，每个图像对中两幅图像的顺序是随机的。以随机的顺序演示一系列图像对，观察者只需根据评分表对两组图像进行分级和评分。单激励连续质量量表法是观察者在一段时间内连续对一系列待测图像进行评分，该待测图像序列可以既包含测试序列又包含其对应的基准序列，对图像的评分基于已评分的待测图像与正在评分的待测图像之间的相互比较，最终综合评分分值和评分时间得到待测图像的质量评分[22]。强迫选择量表法和相似度对比法较为类似，前者需要观察者在同时显示的图像对中选择较高质量的图像，后者则需要在同时显示的图像对中依据等级表选择两者的质量差异分数。观察者会根据这些评价方法和标准、按照主观上对于图像质量的感受对测试图像进行打分，之后再计算所有观察者评分的 MOS 和 DMOS 值，以此作为该测试图像的主观评价质量分数[23]。

目前应用最为广泛的媒体体验质量评价数据库都使用单一的 MOS 或 DMOS 值作为评价指标，而用户评分实际上符合某种具体的分布，因此仅用均值来评价媒体体验质量缺乏合理性和精确性。我们尝试寻找一种更加全面通用的方法对于媒体体验质量进行表征。

本文在媒体主观体验质量分数的统计特性上分布规律做了部分研究，基于 LIVE 数据库中的图像重新构建了自己的数据库，对评分数据进行处理分析，寻找图像的评分统计量与评分分布类型、图像失真类型和图像内容之间的关系，本文剩余部分的研究内容和思路如下。

第二章为数据库构建。本章主要阐述主观质量评价数据库的建立过程，依据 ITU-R BT.500-13 建议书，选定评价方法、选择测试素材和图像序列、设计测试系统、以及获取评分数据。第三章为数据处理分析算法。本章整合了所有对结果的分析步骤。首先，进行数据预处理，即筛选观察者；其次，计算各阶统计量；然后，进行常见概率分布的拟合优度检验，包括正态分布、指数分布、广义帕累托分布等；最后，选择几种评分统计量，研究其与评分分布类型、图像失真类型和图像内容之间的关系。

第四章为结果分析。本章针对上一章数据处理后得到的结果进行分析和总结，探讨图像的评分情况与分布类型、图像内容和失真类型之间会否存在显著规律，以及是否存在不同点或共同点。第五章为总结。本章总结分析了新数据库的统计特性，同时对媒体体验质量评价指标后续优化的方向进行了阐述。

2 数据库构建

2.1 主观评价方法

本测试使用的评价方法为单激励连续质量量表法（SSCQE），即显示一个图像序列并进行评分，该图像序列包含测试序列和对应的参考序列。评价方法中测试素材和观察者的选择及其他要素的规定全部参考 ITU-R BT.500-13 建议书中的说明及要求。

(1) 测试素材的选择

在诸多主流媒体体验质量数据库中，本测试选择 LIVE 数据库进行评价，因为 LIVE 数据库的参考图像数量及失真类型数量适中，而失真图像数量较多，因此有更多可供选择的图像，但又不会因为数量过多而导致素材选择的过程过于繁琐。综上，LIVE 数据库更符合实验需求。

因此，本测试最终在 LIVE 数据库中选择了 100 张图像进行测试，包含 28 种图像内容，平均每种图像内容有 3-4 张图像，并对应不同的失真程度。同时，由于 LIVE 数据库共有 5 种失真类型，因此每种失真类型选择了 20 张图像。LIVE 数据库共有 29 种图像内容，28 种用于测试，剩余 1 种用于预训练，帮助测试者了解整体质量分布。本测试选择的部分测试和训练图像如图 1 和图 2 所示。

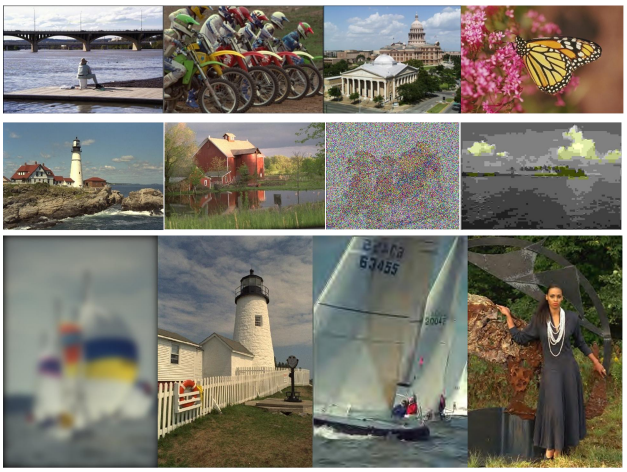

图 1 从 LIVE 数据库中选出的测试图像示例

图 2 从 LIVE 数据库中选出的训练图像

(2) 观察者的选择

一共有 180 名观察者参与测试，其中仅有 2 名观察者是“专家”，熟悉图像处理领域，其余均为“无经验”的观察者。同时，记录了大部分评价人员的特点，包括年龄和职业类别这两个方面的数据。观察者年龄范围跨度较大，在 18-50 岁之间不等，其中绝大部分观察者的年龄大约在 20 和 40 岁。观察者的职业类别也跨度较大，包括学生、财会、人事、销售、自由职业、自主创业等，其中学生占大多数。

2.2 测试系统设计

本测试设计了 2 个测试系统，其中一个为通过 MATLAB GUI 创建的图形用户界面，一个是在网页端的用户界面，提供给观察者多种选择参与的途径，提高了测试的便捷性与参与人员的广泛性。下图为 MATLAB 测试系统的界面示例。

图 3 MATLAB 测试系统界面示例

3 数据处理分析算法

3.1 数据预处理

由于研究[24]表明双激励法和迫选法适用于测试图像损伤范围较小的情况，而单激励法更适用于损伤范围较大的情况，因此结合数据库的图像质量，本课题的图像质量主观评价方法为单激励连续质量量表法（SSCQE），同时参考 BT.500-13 建议书[21]中关于筛选观察者的部分，选择“用于 SSCQE 法的筛选”。整个局部评分反演的检测主要分为两步，先检测并舍弃与均值存在显著偏差的观察者，再次是在不考虑系统偏差的情况下检测并舍弃前后不一致的观察者。采用的是β_2测试：通过计算函数的峰度来确定评分分布是否“正常”，其中峰度为四阶累积量与二阶累积量平方的比值。如果β_2在[2,4]的范围内，那么这一分布被视为“正常”，反之则视为“不正常”。β_2的计算公式如下，其中m_x为x阶累积量。

$$\begin{cases} \beta_{2_{jklr}} = \dfrac{m_4}{(m_2)^2} \\ m_x = \dfrac{1}{N}\sum_{n=1}^{N}\left(u_{njklr} - \overline{u}_{jklr}\right)^x \end{cases}$$

u_{njklr}：观察者n在每一测试配置的每一时间窗口j、测试条件k、序列/图像l、重复r次情况下的评分。

$\overline{u}_{jklr}$：均值。计算公式如下：

$$\overline{u}_{jklr} = \frac{1}{N}\sum_{n=1}^{N} u_{njklr}$$

S_{jklr}：标准差。计算公式如下：

$$S_{jklr} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}\left(u_{njklr} - \overline{u}_{jklr}\right)^2}$$

若$\beta_{2_{jklr}} \in [2,4]$，则

若$u_{njklr} \geq \overline{u}_{jklr} + 2S_{jklr}$，则$P_i = P_i + 1$
若$u_{njklr} \leq \overline{u}_{jklr} + 2S_{jklr}$，则$Q_i = Q_i + 1$

若$\beta_{2_{jklr}} \notin [2,4]$，则

若$u_{njklr} \geq \overline{u}_{jklr} + \sqrt{20}S_{jklr}$，则$P_i = P_i + 1$
若$u_{njklr} \leq \overline{u}_{jklr} - \sqrt{20}S_{jklr}$，则$Q_i = Q_i + 1$

若$\frac{P_i}{J \cdot K \cdot L \cdot R} = \frac{P_i}{100} > 0.2\%$或$\frac{Q_i}{J \cdot K \cdot L \cdot R} = \frac{Q_i}{100} > 0.2\%$，则舍弃观察者$i$。

经过数据预处理的步骤，筛选掉了 17 个观察者，则有效观察者个数为 163。

3.2 计算评分数据的各阶统计量

我们计算每张图像评分数据的常见统计量，以表征评分的数字特征，从而了解该图像主观质量分数的数学性质，进一步推断和评估图像质量。本文计算了每张图像的评分均值、标准差、偏度、峰度和中位数绝对偏差。其中均值是表示集中趋势的统计量，标准差和中位数绝对偏差是表示离散趋势的统计量，偏度和峰度表征的是数据的分布形态。

若一共有N张图像，每张图像有M个打分者，记第m个打分者对第n张图像的评分x_{mn}，第n张图像的评分向量为$\boldsymbol{x_n}$，则这些统计量的计算公式如下：

(1) 均值（Mean）

$$\mu_n = \bar{x}_{\ n} = \frac{1}{M}\sum_{m=1}^{M} x_{mn}$$

(2) 标准差（Standard Deviation）

$$\sigma_n = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}\left(x_{mn} - \bar{x}_{\ n}\right)^2}$$

(3) 偏度（Skewness）

$$\begin{cases} S_{k_n} = \dfrac{\mu_{3_n}}{\sigma_n^3} \\ \mu_{k_n} = \dfrac{1}{M}\sum_{m=1}^{M}\left(x_{mn} - \bar{x}_{\ n}\right)^k \end{cases}$$

(4) 峰度（Kurtosis）

$$K_n = \frac{\mu_{4_n}}{\sigma_n^4}$$

(5) 中位数绝对偏差（Median absolute deviation，MAD）

$$MAD_n = median(|x_{mn} - median(\boldsymbol{x_n})|)$$

3.3 拟合优度检验

研究[15]表明当一个实体（如餐馆、影片、产品等）的专家平均评分为正态分布时，用户平均评分并不是完美的正态分布，而是有较大的偏离，近似服从对数正态分布。结合图像质量主观评价的传统指标 MOS 值，考虑在图像主观质量分数上验证这一结论并做进一步的研究。

3.3.1 绘制直方图

为了下一步假设检验能够更有针对性地进行，绘制每张图像的评分直方图，观察猜测图像质量的主观评分可能符合哪种分布，并进行初步分类。本文利用 MATLAB，对每张图像绘制组数为 10 和 20 的直方图，从粗略和细致两个角度来进行观察分析。观察之后发现主要可分为 4 种有显著规律的直方图，分别是正态分布、左指数分布、右指数分布和广义帕累托分布，其中左指数分布为常见的左偏的指数分布，右指数分布定义为与左指数分布关于$x = 50$对称的右偏的指数分布。图 4-7 分别是猜测符合正态分布，左指数分布，右指数分布和广义帕累托分布的直方图示例。

(1) 正态分布

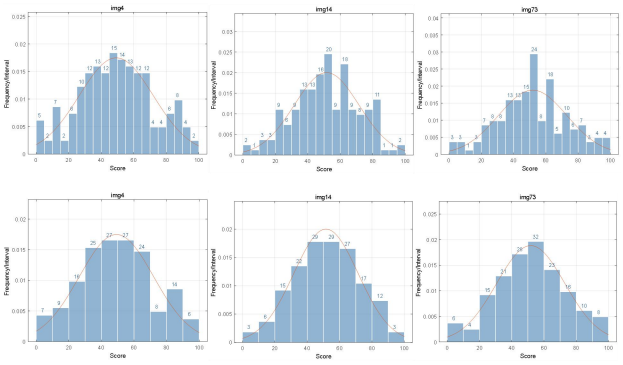

图 4 猜测符合正态分布的图像直方图示例

(2) 左指数分布

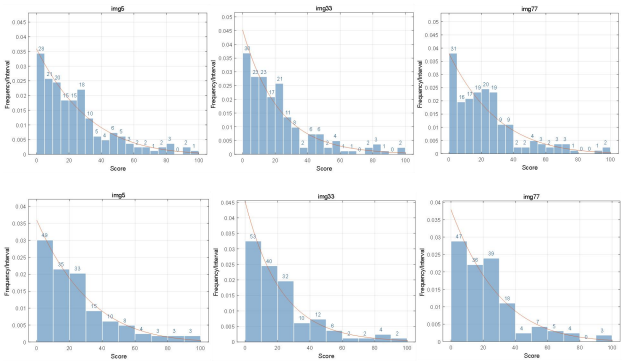

图 5 猜测符合左指数分布的图像直方图示例

(3) 右指数分布

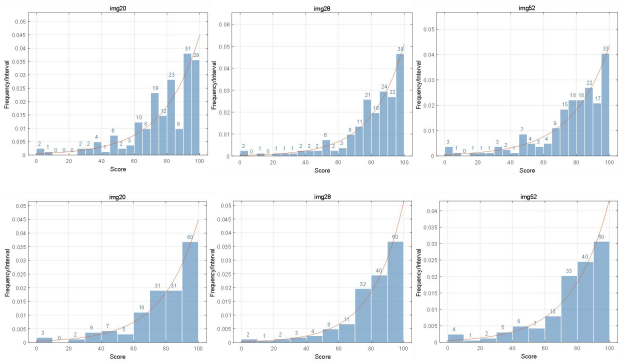

图 6 猜测符合右指数分布的图像直方图示例

(4) 广义帕累托分布

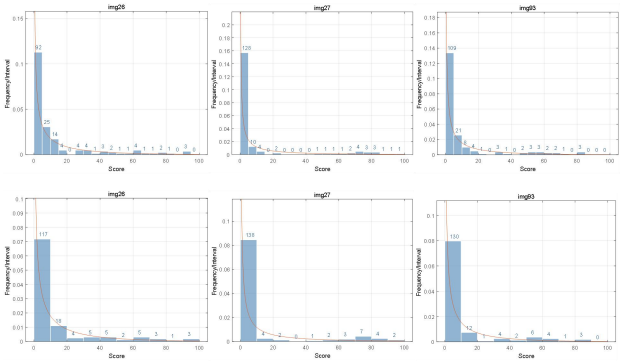

图 7 猜测符合广义帕累托分布的图像直方图示例

3.3.2 假设检验

(1) 剔除异常值

观察所有图像的直方图可以发现，其中有 12 张图像对应的评分分布具有极高的相似性，都近似“广义帕累托分布”，因此暂且将这一类图像的直方图分布定义为“左极值分布”，观察这些失真图像本身及其对应的评分情况，可以发现绝大多数观察者对此类图像质量具有共识，都认为该图像的质量属于“劣”或“差”，而少数观察者认为该图像的质量为“优良”，人数在 8-19 之间不等，因此将这一部分观察者定义为对于某一张左极值图像的异常人群，在进行假设检验时将其单独处理，只考虑主流评分的分布拟合。

(2) 拟合检验

提出零假设H_0：假设主观质量分数符合 6 种分布，即正态分布、左指数分布、右指数分布、半正态分布、广义帕累托分布和伽马分布。对每张图像i、每一种分布q，从$i = 1$、$q = 1$至$i = 100$、$q = 6$，拟合检验的算法如下：

首先，进行参数估计。采用最大似然估计法和矩估计法求解拟合分布未知参数的估计量，本文使用 MATLAB 中的 mle 函数估计每张图像的评分数据服从某种分布的参数值，显著性水平取$\alpha = 0.05$。

其次，计算理论分布每个 bin 的概率。将上一步得到的参数向量记为ψ，则理论分布的累积分布函数可记为$F(x, \psi)$。评分极差为$R = 100$，组数为$K = 20$，则组距为$I = \frac{R}{K} = 5$，设第$k(k = 1,2, ..., n)$个 bin 的概率为P_k，则：

$$P_k = \begin{cases} F(kI, \psi) - F(-\infty, \psi), k = 1 \\ F(kI, \psi) - F\big((k-1)I, \psi\big), k = 2,3, ..., K-1 \\ F(+\infty, \psi) - F\big((k-1)I, \psi\big), k = K \end{cases}$$

最后，进行卡方检验。卡方检验可以统计样本

的实际观测值与理论推断值之间的偏离程度。已知样本容量为$N=163$，记样本的频数分布直方图中每个 bin 对应的频数为f_0，理论分布每个 bin 对应的频数为$f_e=P_k\cdot N$，则卡方统计量的计算公式为

$$\chi^2=\sum_{k=1}^{K}\frac{(f_0-f_e)^2}{f_e}$$

同时χ^2临界值χ_α^2可由卡方分布的逆累积分布函数计算可得。卡方分布的逆累积分布函数记为$G(1-\alpha,v)$，其中$1-\alpha$为置信水平，v为自由度。当参数向量ψ所含元素个数为m时，$v=K-1-m$，$\chi_\alpha^2=G(0.95,v)$。比较χ^2和χ_α^2，若$\chi^2<\chi_\alpha^2$，则接受假设H_0，即符合理论分布，令$H_0=0$；反之则拒绝，令$H_0=1$。

3.4 评分统计量与分布类型的关系

为了验证仅用平均主观质量分数，即 MOS，来表示媒体体验质量有失合理性，在假设检验结果的基础上，寻找评分均值和评分方差与分布类型的关系。

所有图像的评分情况可分为 7 种分布，即正态、左指数（伽马）、右指数、半正态、广义帕累托、伽马和其他分布。其中，“左指数（伽马）”指的是既符合伽马分布又符合左指数分布的评分分布，因为指数分布是伽马分布的特殊情况，两者的联系如下所示；“其他分布”指的是不符合前 6 种分布且没有显著规律的评分分布。

伽马分布的概率密度函数：$f(x,\beta,\alpha)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$，$x>0$。$\alpha=1$ 时为指数分布。

指数分布的概率密度函数：$f(x,\lambda)=\lambda e^{-\lambda x}$，$x>0$。

本文中数据库一共有$N=100$ 张测试图像，令只符合第$q(q=1,2,\dots,7)$种分布的图像数量为N_q，则N_q符合约束$\sum_{q=1}^{7}N_q=1$，这些图像中图像$i(i=1,2,\dots,N_q)$对应的评分统计量分别记为均值μ_{q_i}、评分方差$\sigma_{q_i}^2$、评分偏度S_{kq_i}、评分峰度K_{nq_i}、评分中位数绝对偏差MAD_{q_i}。

对于前 6 种分布N_q的定义和计算方式，进一步的解释说明如下：如果图像i符合单种分布q，则只需对N_q进行循环累加，即$N_q=N_q+1$；如果图像i符合多种分布，则选取其中 P 值最大的分布作为拟合结果最优的分布。由此可计算出满足上述要求的N_q，最后得到的结果如下表 1 所示。

表 1 评分分布类型与对应的图像数量N_q

分布类型	正态	左指数	右指数	半正态
图像数量	23	8	18	6
分布类型	广义帕累托		伽马	其他
图像数量	4		9	32

为研究评分均值和方差、偏度、峰度与评分分布类型的关系，绘制散点图，令横轴为评分均值，纵轴为评分方差或偏度或峰度，不同分布类型的点使用不同的颜色。

3.5 评分统计量与失真类型的关系

本文的测试图像包含 5 种失真类型，分别是白噪声（WN）、高斯模糊（GB），JPEG 压缩（JPEG）和 JPEG2000 压缩（JP2K）和快速衰落（FF）。为研究主观质量分数的整体分布和数字特征是否与失真类型有关，例如在相同分布类型和评分均值的情况下，观察者是否更容易对 JPEG 类型的图像有共识。因此绘制了评分均值、方差和中位数绝对偏差与分布类型的散点图。

令 JP2K、JPEG、WN、GB 和 FF 这 5 种失真类型分别对应数字$p=1-5$，测试图像数量为$N=100$，图像$i(i=1,2,\dots,N)$对应的评分均值为μ_i、评分方差为σ_i^2、评分中位数绝对偏差为MAD_i，以及失真类型为p_i。

为研究评分均值和方差与图像失真类型的关系，绘制横轴为评分均值、纵轴为评分方差的散点图，不同失真类型的点使用不同的颜色。

3.6 评分统计量与图像内容的关系

为研究观察者的评分是否与特定种类的图像内容有潜在联系，例如评分均值相同时，观察者是否更容易对令人心旷神怡的自然风景给出更相近的分数。本文按照图像内容对测试图像进行分类，并绘制散点图进行观察分析。

首先，对图像内容分类。依据 LIVE 数据库中参考图像的命名以及一些图片和摄影网站中对图片的分类规则，可将 LIVE 数据库的图像内容分为 7 种：自然风光、建筑环境、交通运输、运动体育、人物肖像、物品物件和动物图片。自然风光对应

LIVE 库中的“大海”和“溪流”。建筑环境对应“房子”、“灯塔”、“喷泉”、“教堂和国会大厦”、“墓地”、“城市风光”、“雕塑”。交通运输对应“飞机”、“帆船”。运动体育对应“皮划艇”、“越野摩托”。人物肖像对应“女人”、“跳舞”。物品物件对应“帽子”、“玩偶”。动物图片对应“蝴蝶”、“鹦鹉”。

令以上 7 种图像内容分别对应数字$c = 1 - 7$，测试图像数量为$N = 100$，图像$i(i = 1,2, \ldots, N)$对应的评分均值为μ_i、评分方差为σ_i^2、评分中位数绝对偏差为MAD_i，以及图像内容为c_i。绘制横轴为评分均值、纵轴为评分方差的散点图，不同图像内容的点使用不同的颜色，则根据上述参数可得图像i所对应的散点坐标为(μ_i, σ_i^2)。

4 结果分析

4.1 拟合优度检验结果

经过假设检验之后，得到每张图像拟合每种分布的 H_0、P 值和估计的理论分布的参数值。其中符合正态分布图像数量最多，其次是伽马分布和右指数分布，符合 2 种及以上分布的图像数量为 15，而 6 种分布都不符合的图像有 32 张。

根据拟合优度检验得到的结果，下面给出部分符合上述各种分布的图像及其评分直方图示例，如图 8-13，并把不符合这 6 种分布的评分分布定义为其他分布，也给出相应的例子，如图 14。经过观察，可以发现分布类型相同的图像的质量从肉眼直观上得到的感受相似，且相同分布类型的图像的失真类型各不相同。

(1) 正态分布

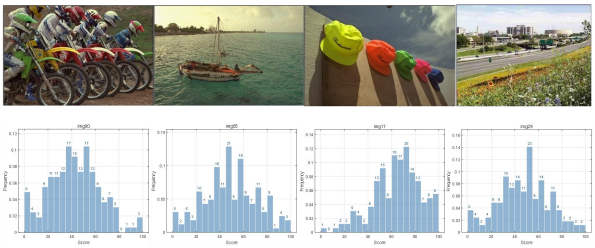

图 8 正态分布部分测试图像及对应的直方图

(2) 左指数分布

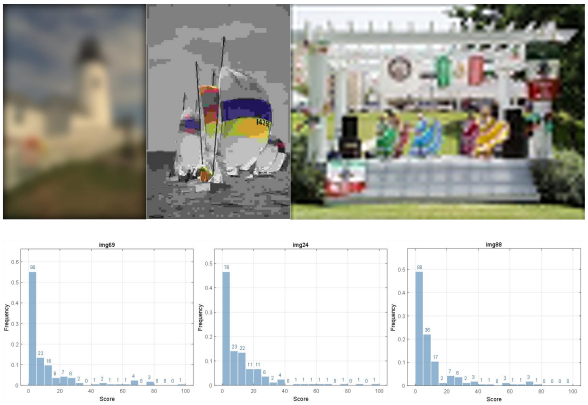

图 9 左指数分布部分测试图像及对应的直方图

(3) 右指数分布

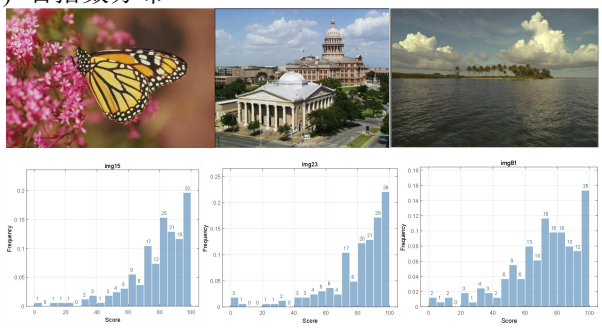

图 10 右指数分布部分测试图像及对应的直方图

(4) 半正态分布

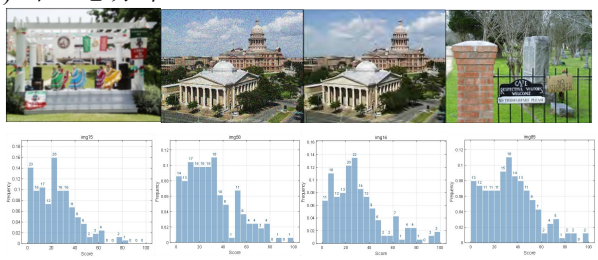

图 11 半正态分布部分测试图像及对应的直方图

(5) 广义帕累托分布

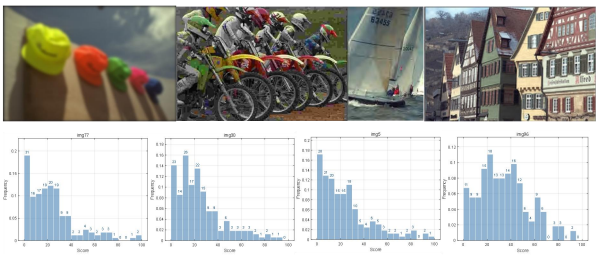

图 12 广义帕累托分布部分测试图像及对应的直方图

(6) 伽马分布

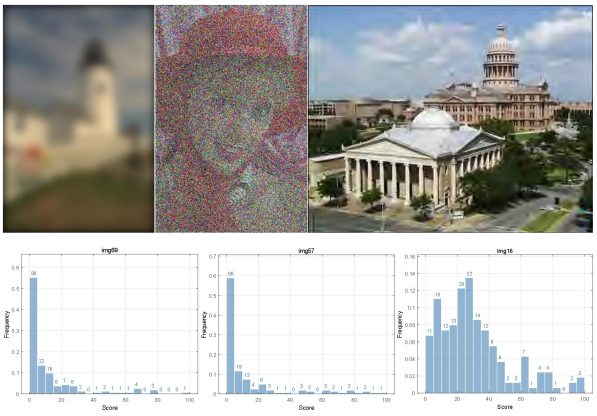

图 13 伽马分布部分测试图像及对应的直方图

(7) 其他分布

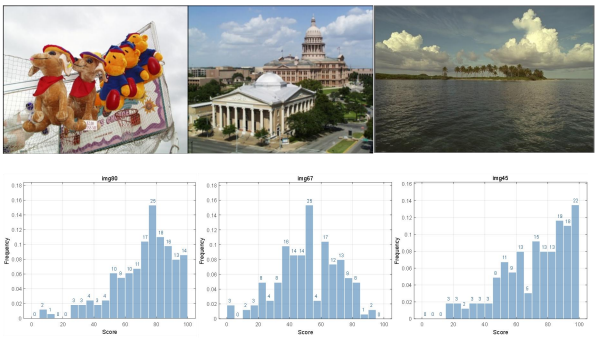

图 14 其他分布部分测试图及对应的直方图

4.2 评分统计量与分布类型的关系的结果分析

(1) 评分均值与失真类型的关系

我们将评分均值和分布类型的关系绘制了散点图，如图 15，可以发现正态、左指数、右指数、半正态、广义帕累托和伽马分布这 6 种分布类型与评分均值有显著关系，而其他分布与均值无关。和预期相同，评分符合正态分布的图像的评分均值处于[0,100]范围的中间部分，符合左指数分布的图像的评分均值偏小，符合右指数分布的图像的评分均值偏大，符合半正态和广义帕累托分布的图像的评分均值处于左指数分布与正态分布之间，最后，符合伽马分布的图像的评分均值均小于正态分布。然而，符合其他分布的图像的评分均值无明显规律，均值覆盖范围较大，其中大部分的评分均值集中在(60,80)之间，与符合右指数分布的图像的评分均值的范围较为接近。各种分布对应的具体的均值范围如下表 2 所示。

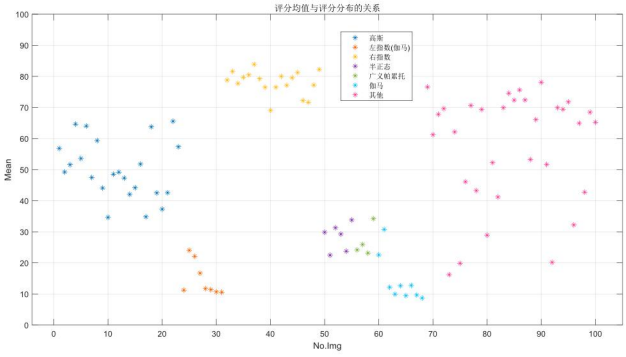

图 15 评分均值与分布类型的关系

表2 分布类型与对应的均值范围

分布类型	正态	左指数	右指数
均值范围	[34.64,65.56]	[10.55,24.04]	[69.08,83.86]
分布类型	半正态+广义帕累托	伽马	其他
均值范围	[22.48,34.22]	[8.703,30.77]	[16.21,78.07]

(2) 评分均值和方差与分布类型的关系

此外，我们也将均值和方差以及分布类型绘制了散点图，如图 16，横轴为均值，纵轴为方差，可以发现评分方差与评分分布类型无显著关系。评分均值近似相等的情况下，图像的评分方差大小各异，且较大的方差和较小的方差对应的图像的评分分布类型无明显规律。在 4.3 节中将具体讨论均值近似相等时方差值相差较大的图像有无相似性或在何处有所区别。

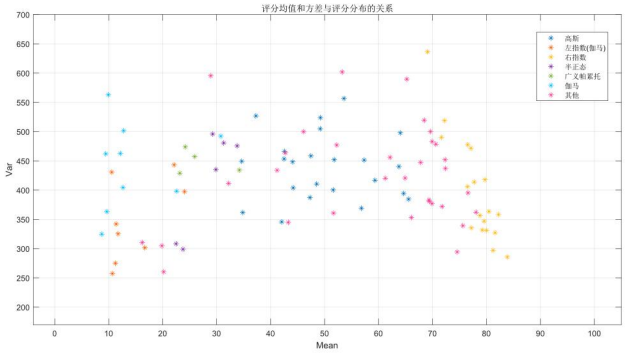

图 16 评分均值和方差与分布类型的关系

(3) 评分均值和偏度与分布类型的关系

我们也将均值，偏度与分布类型绘制了散点图，如图 17，横轴为均值，纵轴为偏度。可以发现均值和偏度近似呈斜率为负的线性关系，由于均值和分布类型存在一定的对应关系，因此评分偏度的大小受评分分布类型的影响。符合左指数分布和伽马分布的图像的评分偏度最大，其次是半正态分布和广义帕累托分布，然后是正态分布，其偏度接近为 0，可构成近似$y = 0$ 的水平线，与理论相符，最后是右指数分布。其他分布中大多数图像的评分偏度处

在正态分布和右指数分布的偏度值之间。具体偏度范围如下表 3 所示。

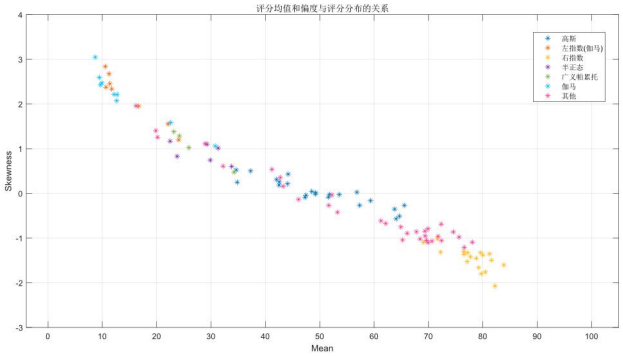

图 17 评分均值和偏度与分布类型的关系

表3 分布类型与对应的偏度范围

分布类型	正态	左指数	右指数
偏度范围	[-0.568,0.525]	[1.198,2.84]	[-2.072,-1.017]
分布类型	半正态+广义帕累托	伽马	其他
偏度范围	[0.474,1.38]	[1.061,3.047]	[-1.21,2.337]

(4) 评分均值和峰度与分布类型的关系

最后，我们将均值和峰度与分布类型绘制了散点图，如图 18，横轴为均值，纵轴为峰度。可以发现评分均值与峰度的关系近似呈开口向上的碗型曲线，同时评分峰度与分布类型有明显的对应关系。符合左指数分布和右指数分布的图像的评分峰度最大，其次是半正态分布和广义帕累托分布，最后是正态分布，其峰度几乎呈水平直线，约等于 3，与理论值相同。其他分布中大多数图像的评分峰度与半正态和广义帕累托分布的评分峰度相同。具体峰度范围如下表 4 所示。

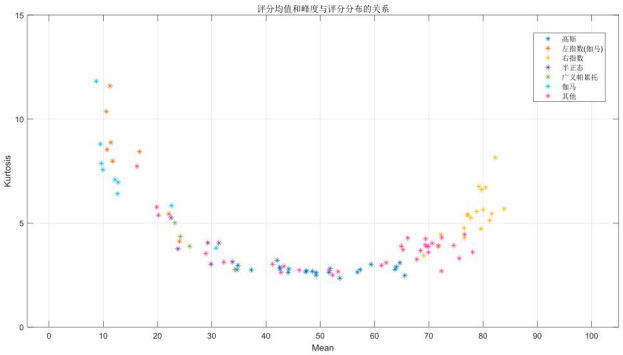

图 18 评分均值和峰度与分布类型的关系

表4 分布类型与对应的峰度范围

分布类型	正态	左指数	右指数
峰度范围	[2.48,3.209]	[4.121,11.59]	[3.436,8.147]
分布类型	半正态+广义帕累托	伽马	其他
峰度范围	[2.769,5.258]	[3.801,11.82]	[2.496,7.731]

4.3 评分均值与方差的关系的结果分析

观察图 16 可以发现，相同评分均值的图像的评分方差大小各异，说明对于质量相近的图像，观察者会因为一些潜在的因素而产生共识或分歧，且无明显规律。本文在评分均值偏小、居中和偏大的情况下，各选择了 3 个数值，比较在评分均值近似相等的情况下（精度约为 1），方差较小和较大的图像及其直方图。

4.3.1 均值偏小

均值偏小的情况下，选择均值分别近似于 10、12、20 的情况。部分示意图如图 19-21 所示。

(1) 均值范围为[9.927,10.68]

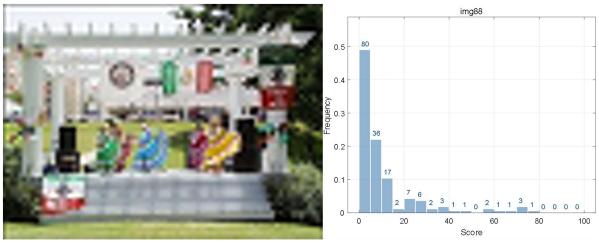

图 19 测试图像 88（快速衰落，均值 10.68，方差 257.4）及其对应的直方图

(2) 均值范围为[11.73,12.75]

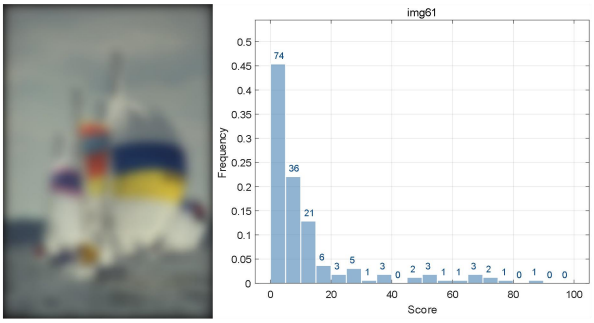

图 20 测试图像 61（高斯模糊，均值 11.73，方差 325.3）及其对应的直方图

(3) 均值范围为[28.9,29.86]

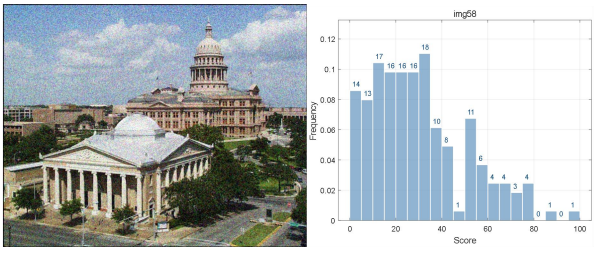

图 21 测试图像 58（白噪声，均值 29.86，方差 435.1）及其对应的直方图

4.3.2 均值居中

均值居中的情况下，选择均值分别近似于 43、52、66 的情况。部分示意图如图 22-24 所示。

(1) 均值范围为[42.57,43.27]

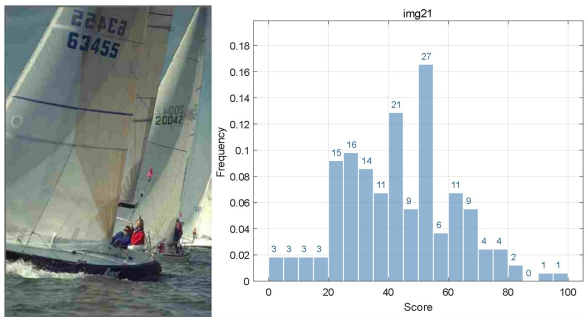

图 22 测试图像 21（jpeg，均值 43.27，方差 344.9）及其对应的直方图

(2) 均值范围为[51.65,53.26]

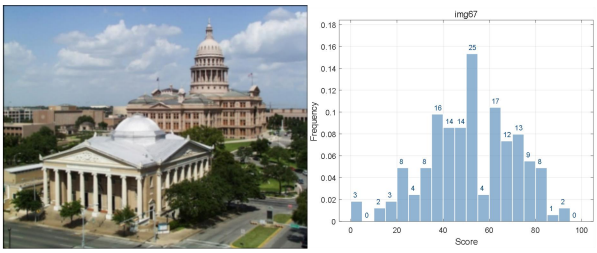

图 23 测试图像 67（高斯模糊，均值 51.65，方差 360.7）及其对应的直方图

(3) 均值范围为[65.23,66.09]

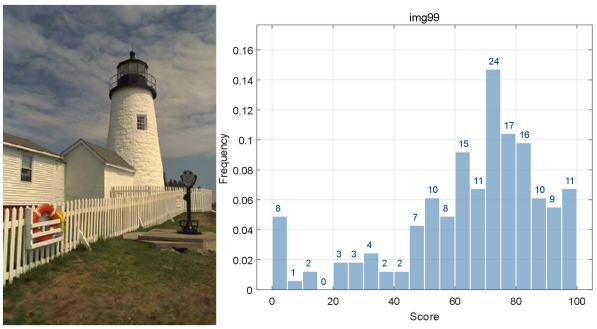

图 24 测试图像 99（快速衰落，均值 65.23，方差 589.6）及其对应的直方图

4.3.3 均值偏大

均值偏大的情况下，选择均值分别近似于 69、72、76 的情况。部分示意图如图 25-27 所示。

(1) 均值范围为[69.08,69.4]

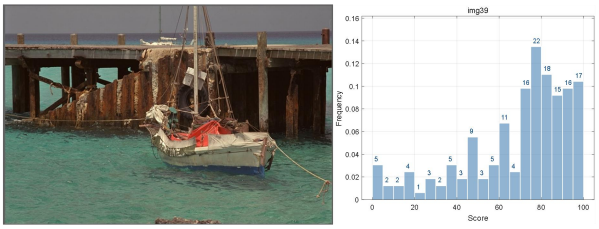

图 25 测试图像 39（jpeg，均值 69.08，方差 636.5）及其对应的直方图

(2) 均值范围为[71.77,72.22]

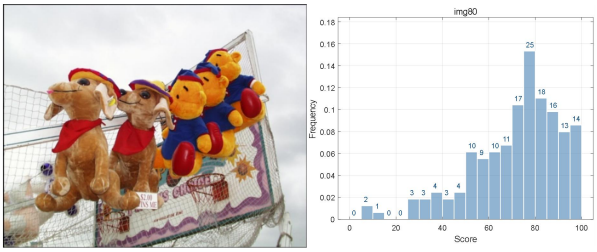

图 26 测试图像 80（高斯模糊，均值 71.77，方差 372）及其对应的直方图

(3) 均值范围为[75.62,76.52]

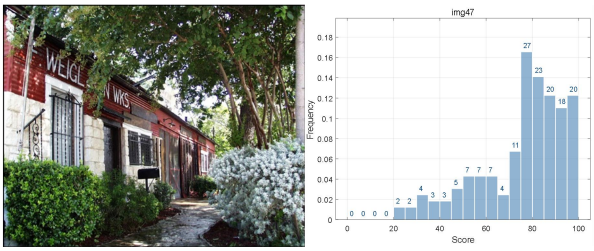

图 27 测试图像 47（白噪声，均值 75.62，方差 339.3）及其对应的直方图

4.4 评分统计量与失真类型的关系的结果分析

实验前猜测图像失真类型为高斯模糊和白噪声时评分方差会相对较小，即观察者对这类图像有较大的共识，然而实验结果与猜测不符。我们绘制了评分均值和方差与失真类型的散点图，如图 28 所示。观察图可以发现整张散点图中各点的色彩杂乱无章，从纵向和横向看，图像的评分均值与失真类型无显著关系，同时，方差偏小、居中、偏大时对应的失真类型也各不相同，说明图像的评分方差和失真类型无紧密联系。

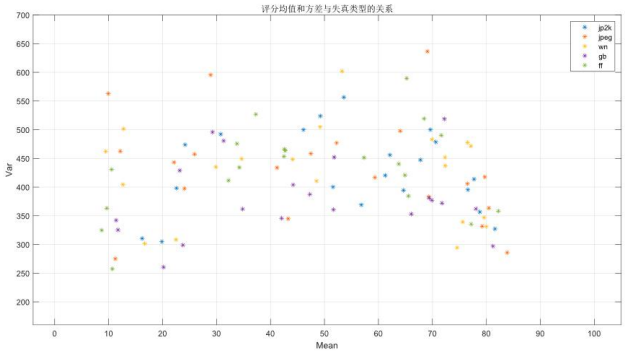

图 28 评分均值和方差与失真类型的关系

4.5 评分统计量与图像内容的关系的结果分析

我们也绘制了评分均值和方差绝对偏差与图像内容的散点图，如图 29 所示。图中的散点颜色错综复杂，观察后可以发现图像的评分均值与图像内容无显著关系，同时，评分方差与图像内容也无明显联系。相同均值或方差的情况下，图像内容各不相同。

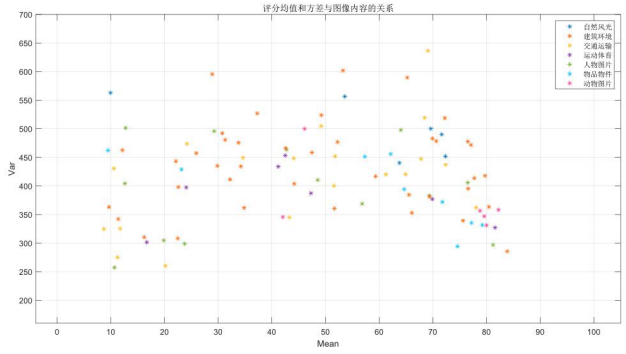

图 29 评分均值和方差绝对偏差与图像内容的关系

5 结论

本文基于 LIVE 数据库中参考图像与失真图像的重新构建了多人打分的媒体体验质量数据库，共有 180 名观察者对 100 张测试图像进行了 0-100 的评分，用于研究媒体主观体验质量中的评分统计和分布规律。

对数据进行预处理后，计算了 5 种经典的常见统计量，并进行了拟合优度检验，包括正态、左指数、右指数、半正态、广义帕累托和伽马分布。最终结果是评分符合正态分布的图像占比约为四分之一，符合左右指数分布的图像占比不到三分之一，同时还有不符合这 6 种分布的图像，约占总数的三分之一。

研究了一些评分统计量（如均值、方差、中位数绝对偏差）与评分分布类型、图像失真类型和图像内容的关系。具体如下：

关于评分统计量与分布类型的关系：图像的评分均值、偏度和峰度均与评分分布类型有较大关联，同时，由于评分均值近似相同的图像的评分方差大小各异，因此评分均值与方差之间无显著联系。

评分均值与分布类型的关系：上述 6 种分布类型均与评分均值相互对应，而其他分布与均值无关。评分符合正态分布的图像的评分均值居中，符合左指数分布的均值偏小，符合右指数分布的均值偏大，符合半正态和广义帕累托分布的均值处于左指数分布与正态分布之间，最后，符合伽马分布的图像的评分均值均小于正态分布。然而，符合其他分布的图像的评分均值无明显规律，均值覆盖范围较大。

评分偏度与分布类型的关系：均值和偏度近似呈斜率为负的线性关系，而评分偏度的大小受评分分布类型的影响。符合左指数分布和伽马分布的图像的评分偏度最大，其次是半正态分布和广义帕累托分布，然后是正态分布，其偏度接近于 0，最后是右指数分布。其他分布中大多数图像的评分偏度处在正态分布和右指数分布的偏度值之间。

评分峰度与分布类型的关系：均值与峰度的关系近似开口向上的碗型曲线，同时评分峰度的大小受分布类型的影响。符合左指数分布和右指数分布的图像的评分峰度最大，其次是半正态分布和广义帕累托分布，最后是正态分布，其峰度几乎呈$y = 3$的水平直线。其他分布中大多数图像的评分峰度与半正态和广义帕累托分布的评分峰度相同。

评分统计量与图像的失真类型及图像内容无明显联系。

参考文献

[1] 马颂德. 图像与视频的浏览与检索[J]. 中国计算机用户, 2000(10): 25-26.

[2] 高敏娟, 党宏社, 魏立力, et al. 基于非局部梯度的图像质量评价算法[J]. 电子与信息学报, 2019, 41(5): 1122-1129.

[3] 贾惠珍. 基于视觉特性和自然场景统计特性的图像质量评价研究[D]. 南京理工大学, 2016.

[4] 余莎. 基于语义嵌入显著性的图像质量评价方法研究[D]. 西安理工大学, 2016.

[5] 张偌雅, 李珍珍. 数字图像质量评价综述[J]. 现代计算机(专业版), 2017(29): 78-81.

[6] Xu Q, Wu Z, Li S, et al. Bridging the gap between objective score and subjective preference in video quality assessment[J], 2010.

[7] R S H, Z W, L C, et al. LIVE Image quality assessment database release 2.

[8] E L, D C. Consumer subjective image quality database, 2009.

[9] Le Callet P, Autrusseau F. Subjective quality assessment IRCCyN/IVC database[J], 2005.

[10] Ponomarenko N, Lukin V, Zelensky A, et al. TID2008 - A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics[J]. Advances of Modern Radioelectronics, 2009, 10: 30-45.

[11] Ponomarenko N, Jin L, Ieremeiev O, et al. Image database TID2013[J]. Image Commun., 2015, 30(C): 57–77.

[12] 王志明. 无参考图像质量评价综述[J]. 自动化学报, 2015, 41(06): 1062-1079.

[13] Jie X, Xing L, Perkis A, et al. On the Properties of Mean Opinion Scores for Quality of Experience Management[C]. IEEE International Symposium on Multimedia, 2012.

[14] Hoβfeld T, Schatz R, Egger S. SOS: The MOS is not enough![C]. 2011 Third International Workshop on Quality of Multimedia Experience, 2011: 131-136.

[15] Dalvi N, Kumar R, Pang B. Para 'normal' activity: On the distribution of average ratings[J]. Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013, 2013: 110-119.

[16] 何南南, 解凯, 李桐, et al. 图像质量评价综述[J]. 北京印刷学院学报, 2017, 25(02): 47-50.

[17] 李婷. 无参考图像质量评价在车牌筛选中的应用[D]. 内蒙古大学, 2018.

[18] Gur D, Rubin D A, Kart B H, et al. Forced choice and ordinal discrete rating assessment of image quality: A comparison[J]. Journal of Digital Imaging, 1997, 10(3): 103-107.

[19] Liang H, Weller D. Comparison-based Image Quality Assessment for Parameter Selection[J], 2016.

[20] Lewandowska Tomaszewska A. Scene reduction for subjective image quality assessment[J]. Journal of Electronic Imaging, 2016, 25(1): 013015.1-013015.13.

[21] ITU-R Recommendation BT.500-13 (2012). Methodology for the subjective assessment of the quality of television pictures. Geneva, Switzerland: ITU.

[22] 王微. 基于多尺度几何分析的图像质量评价算法研究[D]. 汕头大学, 2014.

[23] Mantiuk R K, Tomaszewska A, Mantiuk R. Comparison of Four Subjective Methods for Image Quality Assessment[J]. Computer Graphics Forum, 2012, 31(8).

[24] Streijl R C, Winkler S, Hands D S. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives[J]. Multimedia Systems, 2016, 22(2): 213-227.