大样本图像质量主观评价方法

刘阳，姜润强*，于洪君，陈健
中国科学院长春光学精密机械与物理研究所，吉林 长春 130033

摘要 针对图像质量数据库的主观评价方法存在失真等级少，缺少实验结果分析等问题，提出一种大样本图像质量主观评价方法。该方法基于双激励连续质量量表进行设计，使用简化的 2 级主观评价尺度评价，通过循环积分、最优选择和顺序调整获得样本图像的质量排序，并基于模糊聚类分析的思想将获得的图像次序的概率视为匹配程度，建立样本的模糊相似矩阵。通过指标规格化，建立模糊相似关系、等价关系以及分类、评分等步骤，最终确定图像质量得分。64 级失真图像质量主观评价实验结果表明，图像质量得分能够准确反映可察觉差异的变化，主观评价结果的正确率达到 94%，图像质量得分的标准差介于 0～7，均值为 3.08(百分制)，远低于其他图像质量数据库的水平。所提方法具有较好的准确性和稳定性，适用于图像质量数据库的主观评价和人眼视觉特性研究。

关键词 成像系统；图像质量评价；主观评价；模糊聚类分析；可察觉差异；图像质量数据库；人眼视觉特性
中图分类号 TP391.41 **文献标识码** A **doi**：10.3788/LOP56.131103

Subjective Image Quality Assessment for Large Samples

Liu Yang, Jiang Runqiang*, Yu Hongjun, Chen Jian
Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, Jilin 130033, *China*

Abstract This study presents a novel subjective image quality assessment for large samples to solve existing problems in subjective assessments of image quality databases, such as less distortion levels and insufficient analysis of experimental results. The proposed method is based on a double-stimulus continuous quality scale and employs a simplified, two-level subjective assessment scale. We obtain a quality sequence of sample images by integrating circularly, selecting the best quality, and adjusting the sequence. Then, fuzzy clustering is used to analyze the quality sequence. The probability of image quality sequence in fuzzy clustering analysis is taken as its matching degree, which establishes a fuzzy similarity matrix of samples. We obtain the image quality score by normalizing the probability, establishing the fuzzy similarity relationship, and building a fuzzy equivalence relation, classification, and scoring. We test the subjective assessment for a 64-distortion-level image. The results demonstrate that the image quality scores accurately reflect the variation of just-noticeable difference, assessment accuracy is up to 94%, standard deviation of the image quality scores is from 0 to 7, and the mean value of standard deviation is 3.08 (percentile system), which is much less than the current level of other image quality databases. The proposed method demonstrates high accuracy and stability, and is suitable for subjective assessments of image quality databases and the study of human visual characteristics.

Key words imaging systems; image quality assessment; subjective assessment; fuzzy clustering analysis; just-noticeable difference; image quality databases; human visual characteristics
OCIS codes 110.3000; 110.2960; 100.3008

1 引 言

图像是人工智能获取外界信息的主要来源，图像质量直接影响计算机信息的获取量。在图像采集、传输、存储和显示过程中，因受到离焦、噪声、压缩、丢包或色彩失真等因素的影响，图像质量会存在

收稿日期：2018-12-05；**修回日期**：2019-01-16；**录用日期**：2019-01-29
基金项目：吉林省科技厅 2017 年重大科技招标专项(20170203015GX)
* **E-mail**：jiang_runqiang@sina.com

不同程度的降低。图像质量评价算法(IQA)能精确反映系统的实时状态,实现智能调整,在目标动态追踪、图像压缩、增强以及降噪等技术中应用广泛[1]。

图像质量数据库是研究 IQA 的重要工具,主要由参考图像、失真图像和对应的主观评价得分组成,数据库可以用于标定算法阈值、训练神经网络[2-3]或通过比较算法与主观评价得分的相关系数评估算法的性能[4],如 Spearman 等级(SROCC)、Kendall 等级(KROCC)、均方根误差(RMSE)等[5-6]。LIVE(image)[7]、IVC[8]、CSIQ[9]、TID2008[10-11] 和 TID2013[12-13]是目前使用频率较高的几个图像质量数据库[14-16]。随着神经网络在 IQA 中的应用和人眼视觉特性的深入研究,上述图像质量数据库暴露出诸多问题。例如,数据库的主观评价结果与可察觉差异(JND)的联系不够紧密[17];样本容量不足,尤其是多失真因素影响的图像[18];Ma 等[19]采用客观评价模拟主观评价结果,并创建图像质量数据库 Waterloo exploration database,但客观评价得分并不适合作为标准去衡量客观评价算法的性能;Kang 等[20]将图像拆成多块以增加训练集的样本容量,但相机采集的图像多为非均匀性失真,各子模块的得分均值并不能完全作为整体图像的质量得分;基于单激励(SS)[21]、双激励损伤量表(DSIS)、双激励连续质量量表(DSCQS)[22]设计的主观评价实验,失真等级少(大多数数据库的失真等级均为 5),统计方法简单,缺少对主观评价结果的挖掘和分析[23]。

扩大数据库图像的失真等级,对多因素影响的图像质量进行主观评价需要增加样本容量,进而导致主观评价结果误差成倍增加,当图像的质量差异接近 JND 时,这种现象尤为明显,且主观评价实验带有强烈的个人色彩和不确定性,某些图像的评价结果十分模糊,难以摸清规律和确定分数。

针对上述问题,本文提出一种适用于大样本容量的图像质量主观评价方法及评价结果的分析方法。主观评价方法基于 DSCQS 实现,采用简化的 2 级主观评价质量尺度以提高方法的准确性,该方法包括循环积分、最优选择和顺序调整 3 个步骤,将全体样本的质量进行初步分类、排序后,通过最优选择不断优化序列,经顺序调整后获得准确的图像质量排序。主观评价结果的分析方法基于模糊聚类实现,核心思想是将图像获得的排列顺序的概率视为两者的匹配程度,进而建立样本的模糊相似矩阵,实现样本聚类,分析人眼视觉特性的变化规律,最终确定图像质量得分。模糊聚类过程中,质量差异与可察觉差异接近的图像因排列顺序的相似程度较高会迅速聚类,随着阈值 λ 降低,差异接近 JND 的图像逐渐聚类,而其他原因造成的误差对图像的相似度影响较小,因此聚类结果所受影响较小。介于模糊统计量的计算过程十分复杂,本文还提出一种确定最佳阈值 λ 的简易方法。

2 图像质量数据库

图像质量数据库对 IQA 的研究至关重要。标定算法阈值、训练卷积神经网络以及度量评价算法性能等工作都需要数据库的支持,目前使用频率较高的几个开源图像质量数据库如下。

1) LIVE,LIVE 由美国德克萨斯大学奥斯汀分校电气与计算机工程系与心理学系联合建立,共包含 29 幅参考图像,有白噪声失真、高斯模糊失真、快速瑞丽衰减失真、JPEG 和 JPEG2000 失真 5 种类型,每种失真类型有 5 个失真等级,由 161 位测试人员参与评分,每幅图像测试 20~29 次,采用 SS 法设计主观评价实验,得分形式为平均主观得分差异(DMOS)[7]。

2) IVC,IRCCyN/IVC 由法国南特中央理工大学建立,共包含 10 幅参考图像,失真类型包括模糊、色彩失真、LAR 编码、JPEG 压缩和 JPEG2000 压缩 5 种类型,每种失真类型有 4 个失真等级,由 15 位测试人员参与评分,每幅图像测试 15 次,采用 DSIS 法设计主观评价实验,得分形式为 DMOS[8]。

3) CSIQ,CSIQ 由美国俄克拉荷马州立大学的电气与计算机工程学院建立,共包含 30 幅参考图像,有高斯模糊、加性高斯白噪声、加性高斯粉红噪声、整体对比度缩减、JPEG 压缩和 JPEG2000 压缩 6 种类型,每种失真类型包含 4~5 个失真等级,由 25 位测试人员参与评分,每幅图像测试 5~7 次,采用 SS 法设计主观评价实验,得分形式为 DMOS[9]。

4) TID2008,TID2008 由乌克兰国家航空航天大学的信号接收、传输与处理系(N504)建立,共包含 25 幅无损参考图像,有高斯模糊、图像去噪、加性高斯白噪声、空间位置相关噪声、掩模噪声、高频噪声、脉冲噪声、量化噪声、JPEG 压缩、JPEG2000 压缩、JPEG 传输错误、JPEG2000 传输错误、不同强度的局部块失真、强度均值偏移和对比度变化等 17 种类型,每种失真类型包含 4 个失真等级,由 838 位来自不同国家的测试人员参与评分,每幅图像测试 33 次,采用 DSCQS 法设计主观评价实验,得分形式为平均主观得分(MOS)[10-11]。

5) TID2013,TID2013 为 TID2008 的升级版,共包含 25 幅无损参考图像,失真类型在原有基础上加入色饱和度变化、乘性高斯白噪声、舒适噪声、噪声图像压缩损伤、数字图像颜色量化偏差、色散和稀疏采样与重构 7 种失真类型,失真类型增加至 24 种,每种失真类型失真等级增加至 5 级,由 985 位来自不同国家的测试人员参与评分,每幅图像测试 47 次,采用 DSCQS 法设计主观评价实验,得分形式为 MOS[12-13]。

上述 5 种常用图像质量数据库的参数如表 1 所示。主观评价实验参数如表 2 所示。表 2 中 Screen 为主观评价实验使用的显示器类型及尺寸,H_s 为屏幕的高度,BMP 表示位图文件,PNG 表示便携式网络图形文件,CRT 表示阴极射线显像管显示器,LCD 表示液晶显示器。

表 1 常用图像质量数据库

Table 1 Widely used image quality databases

Database	Year	Reference	Distortion	Level	Total	Format	Resolution
LIVE(image)	2006	29	5	5	1011	BMP	≤768×512
IVC	2005	10	10	4	195	BMP	512×512
CSIQ	2010	30	6	4-5	930	PNG	512×512
TID2008	2008	25	17	4	1725	BMP	512×384
TID2013	2013	25	24	5	3025	BMP	512×384

表 2 主观评价实验

Table 2 Experiments of subjective assessment

Database	Method	Score	Subject	Rating	Screen	Distance
LIVE(image)	SS	DMOS	161	20-29	CRT/21″	$2H_s$-$2.5H_s$
IVC	DSIS	DMOS	15	15	CRT/21″	$6H_s$
CSIQ	SS	DMOS	25	5-7	LCD/21″	80 cm
TID2008	DSCQS	MOS	838	33	LCD/19″	$2H_s$-$4H_s$
TID2013	DSCQS	MOS	985	47	LCD&CRT/19″	$2H_s$-$4H_s$

以 TID2013 数据库为例,首先简要介绍数据库主观评价实验流程。TID2013 的参考图像中失真因素有 24 种,失真等级为 5 级,失真图像有 120 幅。所采用主观评价方法类似国际象棋循环积分赛的规则,包括以下步骤:

1) 将 120 幅图像两两一组随机分为 60 组,测试者在观察几秒后选择出两幅图像中质量较好的一幅,被选中的图像+1 分,测试软件界面如图 1 所示;

2) 重复 9 次步骤 1),测试者完成 540 道选择题后,测试结束;

3) 汇总测试结果,去除 2%明显错误的主观评分,取分数平均值为该图像的最终得分。

分析 LIVE、TID 等数据库的主观评价实验可知,对于人眼的识别能力,5 级失真等级跨度较大,样本容量小时,如 TID2013 的高斯模糊,每级降质 1.73δ(δ 为标准偏差),结果远超过 JND;主观评价方法缺少对实验数据的挖掘和分析,MOS 值与 JND 没有直接关联。

3 主观评价及分析方法

3.1 主观评价方法

为确保主观评价实验在样本容量增加及图像质量差异接近 JND 时仍具有较高的准确性,简化 ITU-R BT.500-13 提出的主观评价质量尺度[22],仅给予测试者两种选择,ITU-R BT.500-13 的主观评价质量尺度如表 3 所示,简化后的质量尺度如表 4 所示。

大样本图像质量主观评价方法由以下 3 个步骤组成:

1) 循环积分,将待评价的 m 幅图像随机划分为 $m/2$ 组,测试人员对每组图像进行评价,质量较好的图像记 2 分,如果质量相似各记 1 分,进行 r_1 次步骤 1)后,按图像积分由高至低排序,获得序列 $\boldsymbol{O}_1$;

2) 最优选择,测试者从序列 $\boldsymbol{O}_1$ 当前分数最高的 4 幅图像中选出质量最好的图像,将被选中的图像从序列 $\boldsymbol{O}_1$ 移至序列 $\boldsymbol{O}_2$,重复操作直至序列 $\boldsymbol{O}_2$ 中包含全部图像,然后将序列 $\boldsymbol{O}_2$ 作为下一轮测试的序列 $\boldsymbol{O}_1$,重复该操作 r_2 次;

3) 顺序调整,按序列 $\boldsymbol{O}_2$ 的顺序显示全部测试图像,测试者将排列明显错误的图像调整到合适的位置,确认无误后结束测试,获得序列 $\boldsymbol{O}_3$。

增加 r_1 和 r_2 可以有效提高主观评价实验的准确性,但主观评价实验的流程又不能过于繁琐,本研

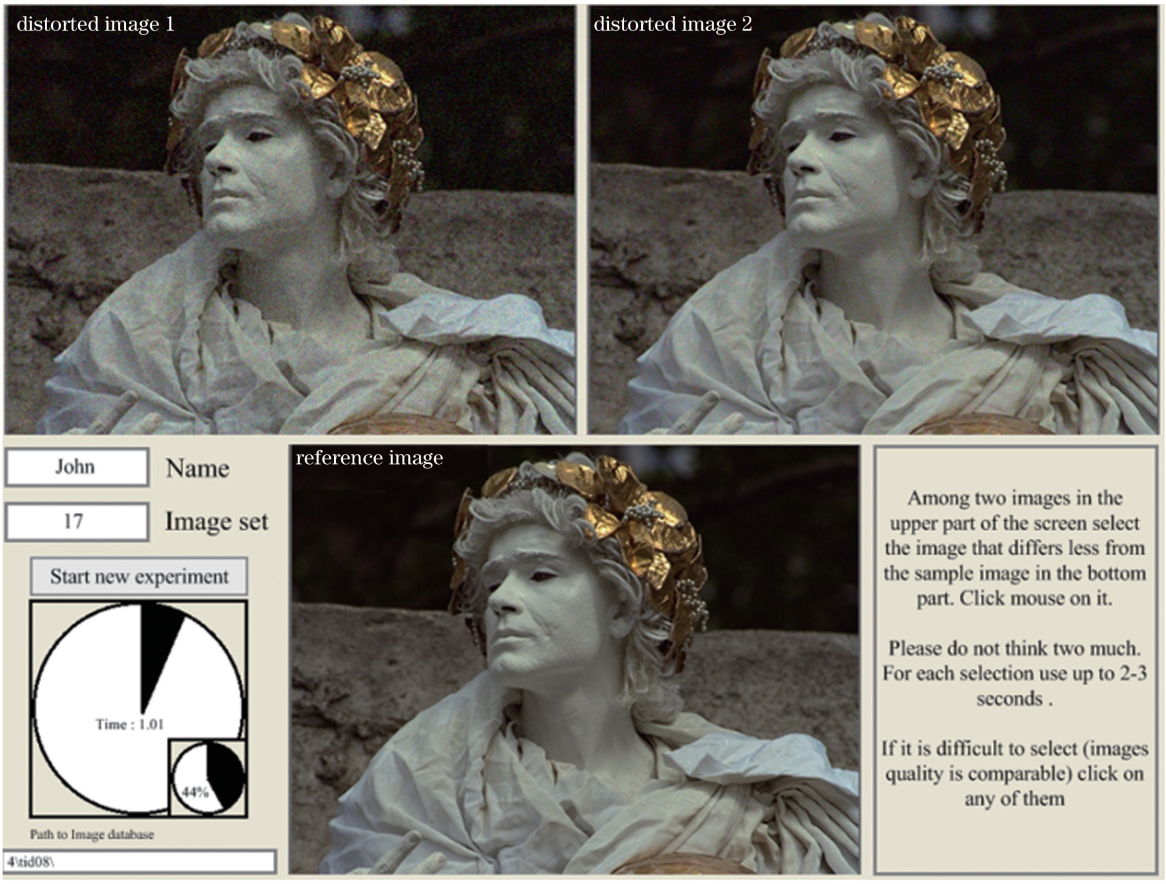

图 1 TID2013 主观评价实验软件界面

Fig. 1 Screenshot of software used in subjective assessment experiments of TID2013

表 3 ITU-R BT.500-13 的主观评价质量尺度

Table 3 Quality scale of subjective assessment for ITU-R BT.500-13

Score	Distortion level	Score	Distortion level
-3	Much worse	+1	Slightly better
−2	Worse	+2	Better
−1	Slightly worse	+3	Much better
0	The same		

表 4 简化后的主观评价质量尺度

Table 4 Simplified quality scale of subjective assessment

Score	Distortion level
+2	Better
+1	The same

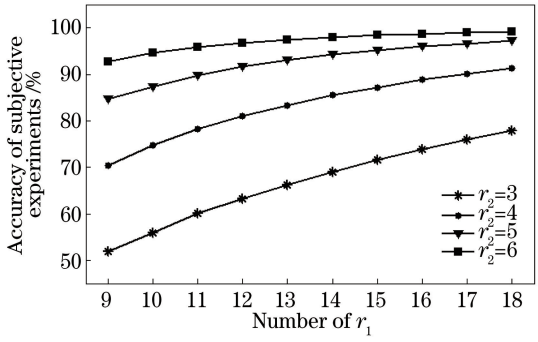

图 2 r_1 和 r_2 对主观评价结果的影响

Fig. 2 Effects of r_1 and r_2 on results of subjective assessment

究方法中 r_1 和 r_2 每增加 1 次，测试题目将分别增加 $m/2$ 和 $m-1$，因此，设计主观评价实验前可通过仿真实验确定样本容量 m 与 r_1、r_2 的最佳关系。假设测试者所有的选择都是正确的，利用 Matlab 软件仿真 r_1 和 r_2 对主观评价结果准确性的影响，图像样本容量 m 为 64，步骤 r_1 和 r_2 的次数分别为 9～18 和 3～6，每组仿真进行 10000 次，所得结果如图 2 所示。

由图 2 可知，$r_1>14$ 且 $r_2>4$ 时，正确率的提高效果减缓，故当 $m=64$ 时，选择 r_1 为 14～16，r_2 为 4～5 相对合理，既能保证较高的正确率，又可以避免实验流程过于冗长。根据前期软件测试的反馈结果可知，步骤 2）容易引起视觉疲劳，进行到第 3 轮时，图像的降质程度已经十分接近人视觉系统阈值，重复测试效果甚微，故当 $m=64$ 时，最终选取 $r_1=16$、$r_2=4$。循环积分、最优选择和顺序调整 3 个实验步骤对应的软件界面分别如图 3～5 所示。

图 3 循环积分测试界面

Fig. 3 Screenshot of software used in cyclical integrating

图 4 最优选择测试界面

Fig. 4 Screenshot of software used in selecting best quality

图 5 顺序调整测试界面

Fig. 5 Screenshot of software used in adjusting sequence

3.2 分析方法

取得多组图像质量的主观评价排序后，基于模糊聚类对主观评价结果进行分析和打分，包括指标规格化、建立模糊相似关系、建立模糊等价关系、分类和打分 5 个步骤。

1）指标规格化，研究对象为主观评价结果，共 m 幅图像 n 次测试，记为 $\boldsymbol{U}=[\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_m]$，其中 $\boldsymbol{u}_i$ 为第 i 次测试获得的图像序列 $\boldsymbol{O}_3$，$\boldsymbol{u}_i=[s_{i1},s_{i2},\cdots,s_{in'}]^{\mathrm{T}}$，$s_{ij}$ 为第 j 次测试第 i 幅图像的序号，求解第 i 幅图像排序为 k 的概率，有

$$p_{ik}=\frac{N(ik)}{m},\tag{1}$$

式中：$N(ik)$ 为主观评价实验中第 i 幅图像排序为 k 的次数，概率矩阵 $\boldsymbol{P}$ 为

$$\boldsymbol{P}=\begin{bmatrix}p_{11} & p_{12} & \cdots & p_{1m}\\ p_{21} & p_{22} & \cdots & p_{2m}\\ \vdots & \vdots & & \vdots\\ p_{m1} & p_{m2} & \cdots & p_{mm}\end{bmatrix}。\tag{2}$$

对矩阵 $\boldsymbol{P}$ 进行规格化处理，可得

$$x_{ik}=\frac{p_{ik}-p_{k\min}}{p_{k\max}-p_{k\min}},\tag{3}$$

式中：$p_{k\max}$ 和 $p_{k\min}$ 为第 k 列元素中的极大值和极小值。规格化后的矩阵 $\boldsymbol{X}$ 可以描述第 i 幅图像与排序 k 的匹配程度，即

$$\boldsymbol{X}=\begin{bmatrix}x_{11} & x_{12} & \cdots & x_{1m}\\ x_{21} & x_{22} & \cdots & x_{2m}\\ \vdots & \vdots & & \vdots\\ x_{m1} & x_{m2} & \cdots & x_{mm}\end{bmatrix}。\tag{4}$$

2）建立模糊相似关系，通过数量积法建立 $\boldsymbol{u}_i$ 与 $\boldsymbol{u}_j$ 的相似关系，相似系数 r_{ij} 为

$$r_{ij}=\begin{cases}1, & i=j\\ \dfrac{1}{M}\displaystyle\sum_{k=1}^{m}x_{ik}x_{jk}, & i\neq j\end{cases},\tag{5}$$

式中：M 为大于 $\max(\sum_{k=1}^{m}x_{ik}x_{jk})$ 的最小正整数，利用 $\boldsymbol{R}$ 表示模糊相似矩阵，元素 r_{ij} 表示样本 $\boldsymbol{u}_i$ 与样本 $\boldsymbol{u}_j$ 的相似程度，则有

$$\boldsymbol{R}=\begin{bmatrix}r_{11} & r_{12} & \cdots & r_{1m}\\ r_{21} & r_{22} & \cdots & r_{2m}\\ \vdots & \vdots & & \vdots\\ r_{m1} & r_{m2} & \cdots & r_{mm}\end{bmatrix}。\tag{6}$$

3）建立模糊等价关系，$\boldsymbol{R}$ 一般只具有自反性和对称性，需要将其改造为模糊等价矩阵，采用平方法求出 $\boldsymbol{R}$ 的传递闭包 $\hat{\boldsymbol{R}}$，则有

$$r_s(i,j)=\bigvee_{k\in[1,m]}[r(i,k)\wedge r(k,j)],\tag{7}$$

式中：$r_s(i,j)$、$r(i,k)$、$r(k,j)$ 为矩阵 $\boldsymbol{R}$ 中的元素；$\bigvee_{k\in[1,m]}$ 和 $\wedge$ 分别表示取极大值和极小值，由模糊数学相关定理可知 $\boldsymbol{R}^m$ 一定是 $\boldsymbol{R}$ 的传递闭包 $\hat{\boldsymbol{R}}$；

4）分类，给定适当的 λ 值，求 $\hat{\boldsymbol{R}}$ 的截关系矩阵，获得图像的动态聚类图，因模糊统计量的计算过程十分繁琐，本文提出一种确定最佳阈值的简易评判方法，具体过程为如下：

a）大致确定样本图像质量变化趋势，例如，如果失真图像由参考图像降质获得，则随着降质等级的增加，图像质量降低；如果失真图像由相机采集获得，可以先确定一段范围内质量最好和最差的图像，然后确定其他图像的质量变化趋势；

b）观察图像排序结果的平均值 $\bar{u}$，如果几幅图像排序平均值十分接近，则这几幅图像可以以大概率聚类，反之图像不应该聚类；

c）观察图像的动态聚类图，依靠步骤 a）～b）的信息，选择符合标准的几种聚类结果，经过讨论后，从中选择最佳的阈值和对应的聚类结果；

5）打分，根据图像质量的变化趋势及经验，对少数应该被聚类而没有聚类的图像进行归类，确定最终的排列顺序，并赋予图像质量主观评价得分，分数越低表示图像质量越好，每增加 1 分代表图像降低一个可辨识等级。

4 主观评价实验及结果分析

通过观察图像质量得分与可察觉差异的关系，以及主观评价结果一致性和图像质量得分标准差，验证所提方法的性能。实验共有 10 名测试人员参与，每人进行 5 次测试，实验环境条件基本一致，显示器为分辨率 1920 pixel×1080 pixel 的 21 寸 LCD 显示器，观察距离为显示器高度的 2.2 倍。图像的失真类型为点扩展模糊和因相机离焦造成的图像失真[24]，共包含 64 级失真，是其他数据库失真样本容量的 12.8 倍，点扩展模糊的半径 r 为图像失真等级的 5/16，参考图像为国内某厂家无人机，大小为 640 pixel×360 pixel，具体如图 6 所示。

图 6 参考图像

Fig. 6 Reference image

使用不同形状、颜色的点区别主观评价实验不同结果出现的概率，其中“·”“◇”“□”“*”“×”分别表示事件出现概率介于(0.8,1.0]、(0.6,0.8]、(0.4,0.6]、(0.2,0.4]、(0.0,0.2]，统计所有实验结果，如图 7 所示，计算图像质量排序的均值，P 表示概率。

统计图像质量主观评价排序，计算传递闭包并绘制研究样本的动态聚类图，部分动态聚类图结果如图 8 所示。

由于主观评价的相似度极高，所以当 $\lambda=0.94$ 时，图像 59、60 率先聚类，随着阈值 λ 逐渐降低，类似 17～18，27～29 的图像逐渐聚类，而序列中相似度非常低的图像，类似 1～16，17 和 19 始终没有聚类。按最佳聚类的筛选标准，经比较可知，$\lambda=$

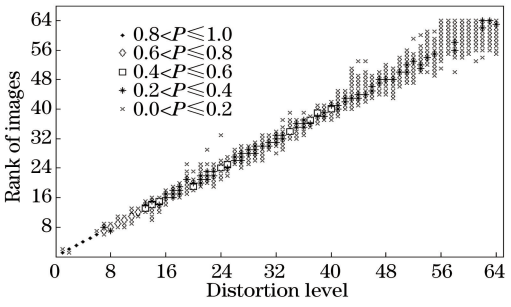

图 7 主观评价实验结果

Fig. 7 Results of subjective assessment experiments

0.4964 时获得图像的最佳分类方式，64 幅图像被划分成 40 类，图像的质量得分如图 9 所示。

通过图像质量主观评价得分，可以估算出 JND 与点扩展模糊半径的对应关系，进而得出：1) 当点扩展模糊半径为 0.31～5.00 时，对应图像 1～16 始终没有聚类，对应的 JND 为 0～+0.31(仅考虑点模糊半径增加的情况)；2) 随着点扩展模糊半径的增加，图像 17～57 对应的 JND 逐渐达到 +0.62、+0.93；(3) 点扩展模糊半径为 17.5 时，图像 58 对应的 JND 超过 +2.5。

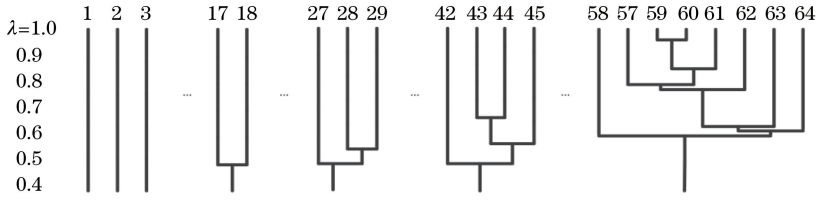

图 8 图像动态聚类图

Fig. 8 Dynamic clustering diagram of images

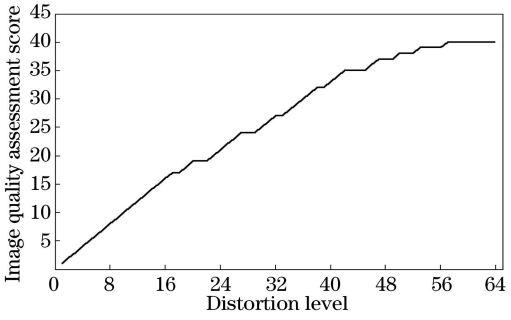

图 9 图像质量主观评价得分

Fig. 9 Subjective assessment scores of image quality

4.1 图像质量得分与 JND

为验证以上结论，基于 DSCQS 法设计 JND 实验，使用表 4 中的绝对评价尺度，参考 Rony Ferzli 的研究中使用的可察觉模糊(JNB)测试方法[25]，测试软件界面如图 10 所示。

令待测图像与降质图像在同一位置显示，每隔 0.5 s 切换一次，由测试人员判断图像清晰程度是否发生变化，如果感觉到图像质量发生变化，则记录数据，指出存在差异的位置并简单描述差异；反之，改变点扩展模糊半径的大小，重新进行测试。两幅图

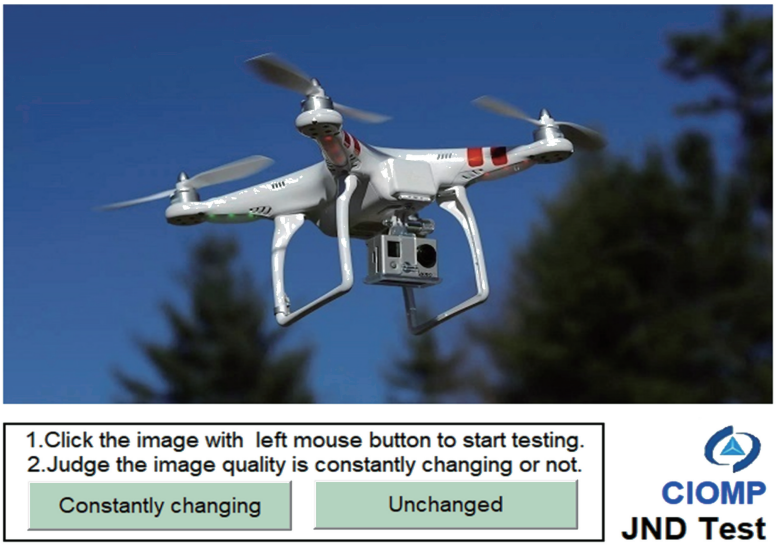

图 10 可察觉差异测试软件界面

Fig. 10 Screenshot of software used in test of just-noticeable difference

像轮流切换以达到增加测试者印象的目的，实验过程中发现逐渐降低图像的质量，测试者的反映会出现“延迟”，结果准确性较差，故选择一幅待测图像和一幅失真图像来回切换；此外，本文提出的图像主观评价测试方法及软件和 JND 测试方法及软件，都尽可能地隐藏每道题目的意图，以避免测试者揣测出测试方法而影响主观评价结果。

为提高实验结果的准确性，JND 测试实验中明确给出图像可察觉差异的定义，即测试者仅能从某些图像的微小细节中隐约感觉到图像的清晰程度发生变化，例如，图 11 中的黑色进气孔栅栏(①、②)，相机侧面的接口和位置(③)，黑色镜头(④)等位置，这些差异需要测试者非常仔细地观察才能发现。定义测试者能够感觉到的整幅图像都出现质量降低的差异为明显变化。

图 11 容易被感知到质量变化的区域

Fig. 11 Areas that are easily perceived to change in quality

选择 9 幅图像进行可察觉差异测试，共有 5 名科研人员参与测试，每人进行 10 次测评，最终测试结果的平均值如表 5 所示。

表 5 平均 JND

Table 5 Mean JND

Distortion level	1	3	5	17	24	32	40	48	56
Point spread blur radius	0.31	0.94	1.56	5.31	7.50	10.00	12.50	15.00	17.50
Just noticeable difference	+0.33	+0.28	+0.34	+0.51	+0.66	+0.97	+1.52	+1.93	+3.35
Obvious difference	+0.69	+0.75	+0.84	+1.86	+2.05	+2.66	+2.84	+3.67	+3.96

降质等级 1、3、5 的图像，JND 分别在待测图像基础上+0.33、+0.28、+0.34，与降质系数比较接近，图像 1～16 的质量得分与质量等级一一对应，结果与结论 1)一致；降质等级 17、24 的图像，JND 系数分别为+0.51、+0.66，数值接近 2 倍降质系数(0.3125)，降质等级 17～25 存在2～3 幅图像质量得分相同的情况，32、40、48 的 JND 系数与图像的聚类情况也基本吻合，结果与结论 2)一致；降质等级 56 的图像，JND 系数为+3.35，降质等级 56～64 的图像均不可识别，图像被聚为一类，结果与结论 3)一致。

降质等级 1、3、5 的图像，人眼可识别的明显变化分别在待测图像基础上模糊+0.69、+0.75、+0.84，接近 2～3 倍降质系数，在最初的图像排序结果中，降质等级 1 的图像排序为 1 和 2 的概率分别为 0.92、0.08，降质等级 3、5 的图像排序为 3、5 的概率为 1；降质等级 17 的图像，明显变化对应+1.86，图像排序为 14～18 的概率分别为 0.02、0.06、0.22、0.58 和 0.12；其他降质等级的图像的得分均处于图像明显变化范围内。

JND 测试实验结果表明，初始的评价结果均处于图像的明显变化范围内，图像质量得分曲线与 JND 的变化规律基本吻合，所提方法能够识别图像的微弱差异，很少出现明显的错误评价，图像质量得分曲线准确反映了可察觉差异的变化。

4.2 主观评价结果一致性

因待测试的样本图像 17～64 中存在人眼不可识别的图像质量差异，所以观察主观评价结果的一致性和比较图像得分标准差时，以图像 1～16 的数据为准。由图 9 的图像质量得分可知，图像 1～16 未被聚类和调整顺序，可以将图像的降质等级作为该图像的质量得分真值。将每位测试人员的 5 次主观评价结果作为一组，计算主观评价均值与图像质量得分真值的欧式距离，比较差异、观察实验结果的一致性，欧式距离的求解方法为

$$L_n=\sqrt{(\bar{s}_{n1}-1)^2+(\bar{s}_{n2}-2)^2+\cdots+(\bar{s}_{n16}-16)^2},\tag{8}$$

式中：L_n 为第 n 位测试人员主观评价结果与真值的欧氏距离；$\bar{s}_{nj}$ 为测试人员的主观评价结果，所得欧式距离结果如图 12 所示。

按(8)式的计算方法，主观评价过程中，每出现一次错误的主观评价欧氏距离至少增加$\sqrt{2}$，而该组数据介于[0.8，2.1]，均值仅为 1.17，表明测试者对图像 1～16 的评价差别介于[0.57，1.48]，即主观评价的正确率为 94.8%。实验结果表明，测试者的个人因素对图像质量主观评价得分的影响较小，参与者 5 次主观评价结果均值与真值十分接近，实验具有非常好的一致性和可重复性。

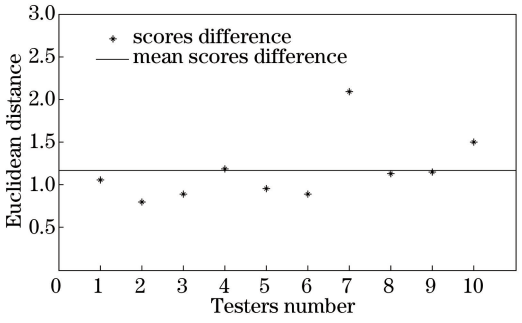

图 12 测试结果与真值的欧式距离

Fig. 12 Euclidean distances between test results and true values

4.3 图像质量得分标准差

计算图像 1～16 主观评价得分的标准差(STD),与其他图像质量数据库的主观评价得分的标准差进行对比。为保证得分标准统一,将得分线性拉伸至满分 100 分,结果如图 13 所示。与其他数据库得分的标准差比较结果如图 14 所示[23](LIVE、TID2008 和 TID2013 数据库未提供主观评价结果的标准差,此处未列入统计)。

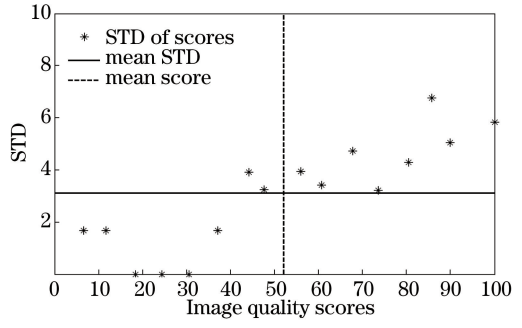

图 13 所提方法所得结果的标准差

Fig. 13 STDs of results obtained by proposed method

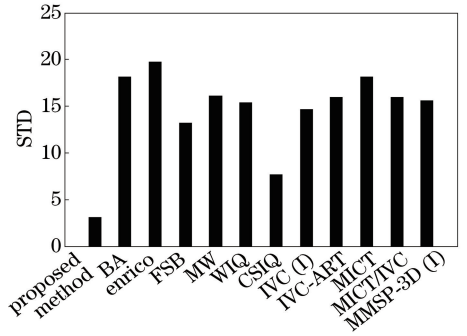

图 14 与其他数据库的标准差比较

Fig. 14 Comparison of STDs between proposed method and other image quality databases

由图 13 可知,本研究图像质量主观评价得分的标准差介于 0～7,均值为 3.08,而其他图像质量数据库的标准差则介于 0～30,标准差均值最低为 7.68(CSIQ 数据库),是本研究标准差均值的 2.49 倍。所提方法获得的图像质量得分标准差远低于其他数据库,主观评价结果具有较好的稳定性。

5 结 论

针对目前图像质量数据库的主观评价方法存在的问题,提出了一种适用于大样本容量的图像质量主观评价方法及评价结果的分析方法。实验结果表明,该方法可应用于大样本容量图像质量评价实验,能够识别图像质量的微弱变化,图像质量得分能够准确反映人眼视觉特性;主观评价结果的准确率达到 94%,百分制下主观评价得分的标准差介于0～7,均值为 3.08,远低于其他数据库的标准差,该方法具有很好的一致性、稳定性和可重复性,适用于图像质量数据库的主观评价实验及人眼视觉特性的研究。

本研究通过设计主观评价实验流程并使用模糊聚类分析评价结果,获得了样本较理想的主观评价得分,随着主观评价实验数量的积累,还会发掘出更合理更高效的实验方法和分析手段。虽然所提方法能够在一定程度上提高主观评价实验的效率,但实验过程仍然十分繁琐,需要耗费大量的人力、物力,仅凭借主观评价获得的图像质量得分难以满足深度学习等研究的需求,因此,采用图像质量数据库,并结合主客观评价的方式(即以对抗网络生成的图像质量得分作为训练集真值,以主观评价实验获得的图像质量得分作为测试集真值),既能够满足研究客观评价算法对样本数量的需求,也可使测试结果更具有强的说服力。

参 考 文 献

[1] Wang Z M. Review of no-reference image quality assessment[J]. Acta Automatica Sinica, 2015, 41(6): 1062-1079.
王志明. 无参考图像质量评价综述[J]. 自动化学报, 2015, 41(6): 1062-1079.

[2] Kim J, Zeng H, Ghadiyaram D, *et al*. Deep convolutional neural models for picture-quality prediction: challenges and solutions to data-driven image quality assessment[J]. IEEE Signal Processing Magazine, 2017, 34(6): 130-141.

[3] Zhang Q B, Zhang X H, Han H W. Optimization of underwater photoelectric image quality based on deep convolutional neural networks [J]. Acta Optica Sinica, 2018, 38(11): 1110004.
张清博, 张晓晖, 韩宏伟. 一种基于深度卷积神经网

络的水下光电图像质量优化方法[J]. 光学学报, 2018, 38(11): 1110004.

[4] Zhang F, Zhang R Y, Li Z Z. Image quality assessment based on symmetry phase congruency [J]. Laser & Optoelectronics Progress, 2017, 54(10): 101003.
张帆, 张偌雅, 李珍珍. 基于对称相位一致性的图像质量评价方法[J]. 激光与光电子学进展, 2017, 54(10): 101003.

[5] Sheikh H R, Sabir M F, Bovik A C. A statistical evaluation of recent full reference image quality assessment algorithms [J]. IEEE Transactions on Image Processing, 2006, 15(11): 3440-3451.

[6] Ma Y M, Chen H Y, Liu G J. General mean pooling strategy for color image quality assessment[J]. Laser & Optoelectronics Progress, 2018, 55(2): 021007.
马月梅, 陈海英, 刘国军. 彩色图像质量评价的广义平均池化策略[J]. 激光与光电子学进展, 2018, 55(2): 021007.

[7] Sheikh H Z, Wang Z, Cormack L, *et al*. LIVE image quality assessment database release 2 [OL]. (2006) [2018-10-24]. http://live.ece.utexas.edu/research/quality/subjective.htm.

[8] Callet P Le, Autrusseau F. Subjective quality assessment IRCCyN/IVC database[OL]. (2005) [2018-10-24]. http://www.irccyn.ec-nantes.fr/ivcdb/.

[9] Larson E C, Chandler D M. Consumer subjective image quality database [OL]. (2009) [2018-10-24]. http://vision.okstate.edu/index.php?loc=csiq.

[10] Tampere image database 2008 TID2008 [OL]. (2008) [2018-10-24]. http://www.ponomarenko.info/tid2008.htm.

[11] Ponomarenko N, Lukin V, Zelensky A, *et al*. TID2008: a database for evaluation of full-reference visual quality assessment metrics [J]. Advances of Modern Radioelectron, 2009, 10: 30-45.

[12] Tampere image database 2013 TID2013 [OL]. (2013) [2018-10-24]. http://www.ponomarenko.info/tid2013.htm.

[13] Ponomarenko N, Jin L N, Ieremeiev O, *et al*. Image database TID2013: peculiarities, results and perspectives [J]. Signal Processing: Image Communication, 2015, 30: 57-77.

[14] Ninassi A, Callet P L, Autrusseau F. Pseudo no reference image quality metric using perceptual data hiding[J]. Proceedings of SPIE, 2006, 6057: 60570G.

[15] Larson E C, Chandler D M. Most apparent distortion: full-reference image quality assessment and the role of strategy [J]. Journal of Electronic Imaging, 2010, 19(1): 011006.

[16] Kundu D, Choi L K, Bovik A C, *et al*. Perceptual quality evaluation of synthetic pictures distorted by compression and transmission[J]. Signal Processing: Image Communication, 2018, 61: 54-72.

[17] Chandler D M. Seven challenges in image quality assessment: past, present, and future research [J]. ISRN Signal Processing, 2013, 2013: 1-53.

[18] Hou C P, Ma T T, Yue G H, *et al*. Multiply-distorted image quality assessment based on high-order phase congruency[J]. Laser & Optoelectronics Progress, 2017, 54(7): 071001.
侯春萍, 马彤彤, 岳广辉, 等. 基于高阶相位一致性的混合失真图像质量评价[J]. 激光与光电子学进展, 2017, 54(7): 071001.

[19] Ma K D, Duanmu Z F, Wu Q B, *et al*. Waterloo exploration database: new challenges for image quality assessment models[J]. IEEE Transactions on Image Processing, 2017, 26(2): 1004-1016.

[20] Kang L, Ye P, Li Y, *et al*. Convolutional neural networks for no-reference image quality assessment [C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 1733-1740.

[21] ITU-T Recommendation P.910. Subjective video quality assessment methods for multimedia applications [EB/OL]. (2008) [2018-10-25]. http://handle.itu.int/11.1002/1000/9317-en?locatt=id:2&auth.

[22] ITU-R Recommendation BT.500-13. Methodology for the subjective assessment of the quality of television pictures [EB/OL]. (2012) [2018-10-25]. https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-13-201201-I!!PDF-E.pdf.

[23] Winkler S. Analysis of public image and video databases for quality assessment[J]. IEEE Journal of Selected Topics in Signal Processing, 2012, 6(6): 616-625.

[24] Hong Y Z, Ren G Q, Sun J, *et al*. Analysis and improvement on sharpness evaluation function of defocused image [J]. Optics and Precision Engineering, 2014, 22(12): 3401-3408.
洪裕珍, 任国强, 孙健, 等. 离焦模糊图像清晰度评价函数的分析与改进[J]. 光学 精密工程, 2014, 22(12): 3401-3408.

[25] Ferzli R, Karam L J. A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB) [J]. IEEE Transactions on Image Processing, 2009, 18(4): 717-728.