静态图像压缩评估方法测评

文义红 杨凯 李波
(北京航空航天大学计算机学院，北京 100191)

摘要 静态图像压缩领域提出了众多图像质量客观评估方法，但对于各种客观评估方法与主观感受相符程度的研究仍有不足。为此，文章设计了一种评价客观评估方法的测评方法：按特定原则建立相对完备的测评图像库，以实际观测得到的主观分数为参考，以客观评估方法的评价分数为测评对象，通过对比同一压缩算法不同压缩倍数、不同压缩算法主观质量相近情况下各种评估方法的表现，以及依据VQEG报告的度量准则所求的度量结果，对客观评估方法的性能进行评价。同时，文章以PSNR和SSIM为例进行了测评，结果表明SSIM各项指标优于PSNR，与现有结论一致。

关键词 静态图像 人眼视觉系统 图像品质 评估方法

1 引言

目前静态图像压缩效果的评估方法主要分为客观方法和主观方法。客观方法主要从总体上反映原始图像和失真图像的灰度差别，以PSNR值为代表。其特点是速度快、稳定性好，但难以反映人眼的视觉特性和主观感知程度，有时甚至与主观印象相悖[1]。主观方法是指判读员按照已规定好的评价准则，对目标图像进行质量评价。主观方法能够真实反映人眼的主观感受，但是容易受到观察者背景知识、观测动机、心理状态和观测环境等诸多因素的影响，因而其稳定性和客观性较差，而且评估过程比较繁琐，难于实施。因此，寻找一种能较好地反映人眼主观感受的客观评估方法对图像压缩效果测评十分重要。

随着人们对人眼视觉系统(Human Visual System，HVS)功能的理解不断深入，各种基于HVS模型的静态像质评估方法应运而生，在实际应用中正逐步取代传统的均方误差(MSE)或峰值信噪比(PSNR)[2]。但是目前很少有针对静态图像压缩中各种评估方法性能的比较实验或者标准。从已有的文献来看，存在的主要问题有：1) 在静态图像压缩领域，新算法不断提出，但针对这些算法的评估研究很少，如文献[3]，其评测全部集中于JPEG或JPEG2000算法，又如微软公司的HDPhoto算法已经成为JPEG－RX候选标准，却很少看到对其性能的评估。2) 已有文献中，绝大多数主观评测过程都使用单刺激评估方法[4]。但实验心理学的传统结果表明进行主观测试时，相对判断比绝对判断更加稳定、准确。在国际电联(ITU－R)BT.500建议的众多评价方法中，也认为双刺激方法因为采用了基准图像，其结果比单刺激方法具有更高的灵敏度和稳定性[5]。3) 大多数测评实验是为了验证某种客观评估方法而设计的，而不是针对某一领域的应用进行比较[3,6-7]。4) 目前绝大部分实验都是使用彩色图像作为图像库，而彩色图像色度空间的三个通道本身互相影响，而压缩算法一般都是对每个通道进行单独压缩，因此使用彩色图像作为图像库判断的结果无法真实地反映压缩算法的性能。

863计划、国家自然科学基金(60775018)，武器装备预研基金项目资助
收稿日期：2009-04-10。收修改稿日期：2009-05-25

为了比较静态图像(主要针对遥感图像)压缩领域的各种客观评估方法与人眼主观感受的符合程度，本文设计了一种评价客观评估方法的测评方法：在采用不同类型图像、不同原理算法、不同压缩倍数所构造的不同失真类型与失真程度的灰度测试图像的基础上，通过对比各种客观评估方法与主观评估方法的符合程度，对客观评估方法的性能进行评价。同时，使用该测评方法对 PSNR 和 SSIM 进行了测评，对比了两种评估方法的性能。

2　测评方法设计

(1) 测评数据建立

测评数据由原始图像和失真图像组成。由于本研究主要针对遥感图像压缩的失真评价，所以原始图像选取了以遥感图像为主的 10 幅 512 像素×512 像素的 256 级灰度图像(全部来自互联网，如图 1)，为保证最终测试数据的普适性，在选择原始图像时，综合考虑了灰度、对比度、纹理复杂度、小目标密集程度、图像分辨率等多种因素。

失真图像由压缩算法压缩后的还原图组成。为保证测评过程对不同类型的压缩失真具有普适性，选择了 JPEG、JPEG2000、FACTRAL、HDPhoto、BPP[8] 共 5 种不同压缩原理的算法，每种算法又进行了不同失真程度的压缩。具体生成方法如下：使用 JPEG2000、BPP 算法，分别对每幅原图进行 4 倍、8 倍、12 倍、16 倍、20 倍压缩并还原；使用 JPEG、HDPhoto、FACTRAL 算法分别依据不同的品质因子或质量等级对每幅原图进行压缩还原，选取与 JPEG2000 算法形成的还原图视觉质量接近的还原图。最终形成 10×5×5 共 250 幅失真图像。

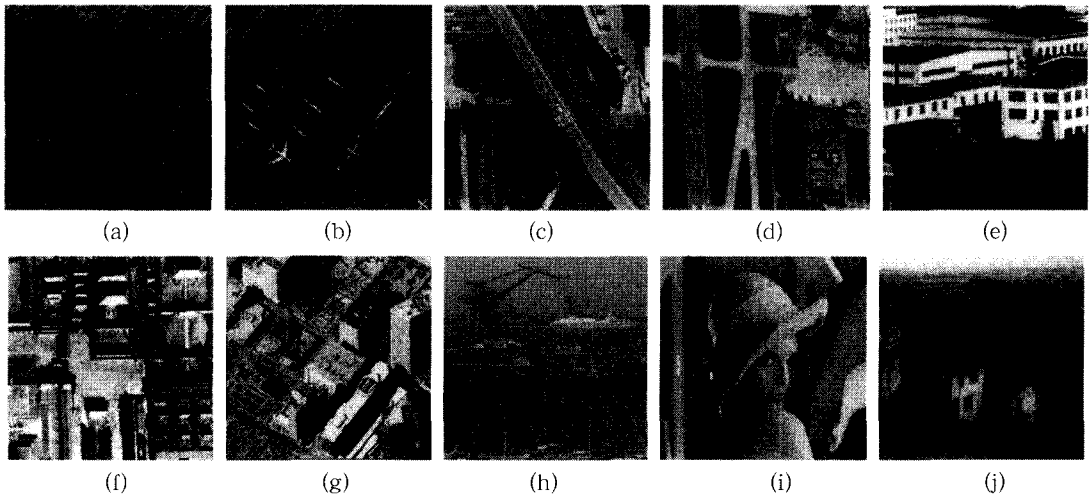


图 1　原始图像

(2) 主观质量评估

主观评估环境：实验场所为内部照明亮度一致的实验室。实验设备为装有 Windows XP SP2 的微机，图像显示软件为 Photoshop，所有显示器均为分辨率设定在 800×600 的 17 寸液晶显示器，都在正常的使用年限之内，并且使用年限相差不超过 2 年。

主观评估方法：对于每幅原始图像，建立一个包含了原始图像及对应所有失真图像的 Photoshop文件，原始图像作为背景，每幅失真图像作为一个图层。首先依据主观质量对不同图层两两比较进行排序，再根据失真图像与原始图像的差别进行打分。打分采用百分制方式，与原图越吻合则分数越高，反之越低。一般的双刺激方法只比较失真图像与原始图像的视觉差别，但本文同

时还比较了失真图像之间的差别，从而更大程度的减少了主观评估中的误差。

观测人员：观测人员有5人，都是图像压缩研究人员，观测者的视力经校正后为正常。观测视距限定在45～50cm。

主观评估分处理：观测人员按照上述评测方法对所有的图像进行评估，每幅失真图像最终主观质量分为5名观察员所打分数的加权平均值。

(3) 客观质量评估

将需要测评的客观评估方法，依据各自的评估算法，对全部测评数据进行评估，并记录评估值。

(4) 测评标准

不同的评估方法由于其算法不同，其评估结果的值域范围各不相同。因此，对数据进行测评比较前，首先要将数据进行拟合。本文的拟合模型使用非线性回归模型与线性回归模型结合的方式：

$$p(x) = b_1/(1+\exp(-b_2\cdot(x-b_3)))+b_4\cdot x+b_5 \tag{1}$$

式中 x 表示实际客观值；p 表示其对应预测值；b_i 为模型参数。实际比较时，使用预测值作为比较对象。

测评时，首先对图像数据在同一压缩算法不同压缩倍数和不同压缩算法主观质量相近两种情况下的测评结果进行分析，统计不同评估方法在图像主观品质不同变化类型下的表现，依据其与主观质量变化所表现的一致性判别其评估性能。同时，参考VQEG[9]报告，选取其中4种度量指标(Pearson线性相关系数(CC)、均方根误差(RMSE)、Spearman秩相关系数(SROCC)、背离率(OR))对实验结果进行分析。其中Pearson线性相关系数计算客观分预测值与实际主观分之间相关系数，该公式反映了主客观分值在该回归模型下的预测相关程度，其值在0～1之间，越大表示主客观分值相关性越强；均方根误差计算客观分预测值与主观分之间的均方误差，它反映了预测值与主观分之间的偏离程度，其值越小说明预测的越准确；秩相关系数计算同一图像在主客观分值序列中排序后的次序位置差异程度[10]，反映模型的预测单调性，其值越小说明主客观分数的趋势越相近；背离率计算与主观分值的差值大于一定阈值的预测值在总样本数的比率，是反映客观评分对主观评分的估计值和主观评分一致性的参量，其值越小表示客观评价模型越好。

最后，通过上述多角度的测评比较，对客观评估方法的性能给出评价。

3 实验及结果分析

作为一种通用的比较静态图像压缩客观评估方法的测评方法，本方法可以评测任何评估方法。为了验证其正确性，本文选用了具有代表性的2种客观质量评估方法PSNR和SSIM[3](Structural SIMilarity)为例进行实验。PSNR是图像压缩领域中最常用的客观品质评估方法。SSIM是目前基于HVS的客观评估方法的代表，现有资料均认为它比PSNR值更符合人眼视觉特性，但是除了文献[3]，并没有其他地方给出验证实验。其核心函数定义为

$$\mathrm{SSIM}(x,y) = [l(x,y)]^{\alpha}[c(x,y)]^{\beta}[s(x,y)]^{\gamma} \tag{2}$$

式中 x、y 分别是原图像信号和失真图像信号；$\mathrm{SSIM}(x,y)$描述图像块失真信号与原始信号之间的相似性，作为失真的最终度量；$l(x,y)$是亮度比较函数；$c(x,y)$是对比度比较函数；$s(x,y)$是结构比较函数；$\alpha>0$，$\beta>0$，$\gamma>0$ 是用于调整权重的参数。三个比较函数定义为

$$l(x,y)=\frac{2\mu_x\mu_y+C_1}{\mu_x^2+\mu_y^2+C_1},\quad c(x,y)=\frac{2\sigma_x\sigma_y+C_2}{\sigma_x^2+\sigma_y^2+C_2},\quad s(x,y)=\frac{\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3} \tag{3}$$

式中　均值 μ_x、μ_y 是亮度的估计；标准差 σ_x，σ_y 是对比度的估计；协方差 σ_{xy} 是结构信息的估计；C 为常数。

根据测评方法设计的过程，对主客观评估方法进行实验，最终得到由人眼观测得到的主观分（表示为 HVS），以及由 SSIM 和 PSNR 计算出来的实际分，图 2 是对应的主客观分的散点分布图。

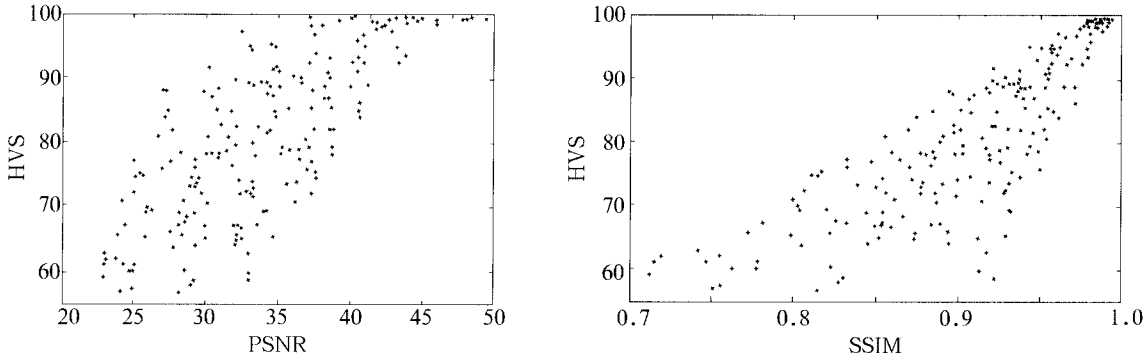
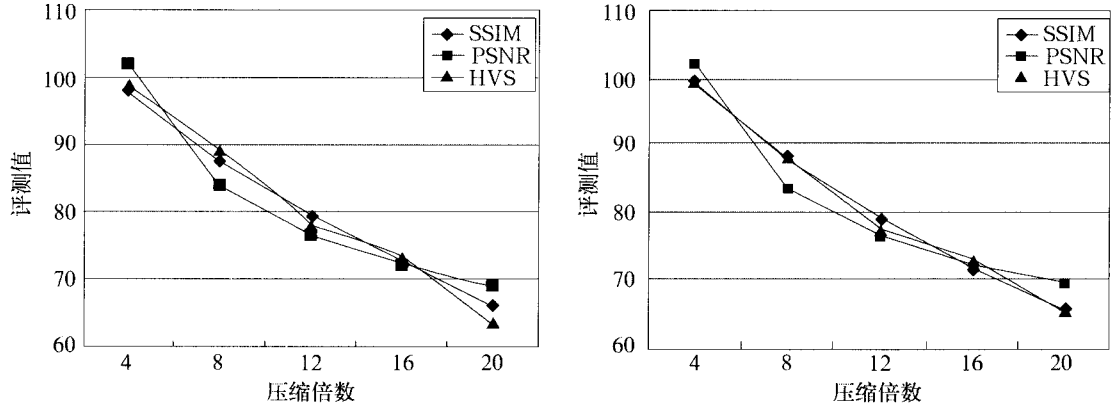


图 2　主客观分的散点图

将不同的评估结果拟合后比较分析，结果如下：

1）对同一压缩算法、不同压缩倍数恢复图评估结果进行比较。由于同一图像在相同压缩算法下，使用不同压缩倍数的恢复像质差别较大，SSIM、PSNR 都与压缩倍数呈单调递减关系。例如图 3(a)是图 1(a)使用 HDPhoto 算法的比较结果，图 3(b)是图 1(a)使用 JPEG200 算法的比较结果，在这种情况下，无论 PSNR 还是 SSIM 都与 HVS 保持较好的一致性。由回归分析可知，SSIM 比 PSNR 更加接近 HVS 曲线，相比之下评估更加精确。



(a) HDPhoto比较结果　　(b) JPEG2000比较结果

图 3　不同压缩倍数恢复图的测评结果

2）对不同压缩算法、主观质量相近恢复图评估结果进行比较。为了便于观察和比较，将同一幅图像在某种压缩倍数下使用不同算法的恢复图像分为一组，这样同组图像较为相似。分析每组数据中与 HVS 趋势一致的点数(见表 1)。在 50 组数据中，SSIM 大于 PSNR 的有 34 组，等于的有 14 组，小于的为 2。SSIM 与 HVS 趋势完全一致的组占 24%，4 点以上一致的占 86%；而 PSNR 与 HVS 趋势 4 点以上一致的占 44%。这说明在主观质量相近的情况下，SSIM 明显优于 PSNR，

前者能够更好地反映压缩失真产生的视觉差别。

表1 主客观指标趋势一致性分析

图像编号	4倍		8倍		12倍		16倍		20倍	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
a	5	4	4	3	4	3	5	3	3	3
b	4	3	5	4	4	3	4	3	4	3
c	4	3	4	3	4	4	4	4	3	2
d	4	3	4	4	5	4	4	3	5	4
e	4	4	4	3	5	3	3	3	4	3
f	3	2	4	3	5	4	4	3	4	4
g	4	5	4	4	5	3	5	4	4	4
h	5	3	4	3	4	3	4	3	4	3
i	3	4	5	3	4	4	5	3	4	4
j	3	3	4	4	4	3	4	3	3	3

3）VQEG度量指标。

依据VQEG的四种度量指标，分别对单幅图像和全部图像进行统计分析，结果如表2。在这些指标中SSIM的Pearson线性相关系数、Spearman秩相关系数都大于PSNR，说明SSIM与主观分之间的相关性和单调性都强于PSNR，并且对于每幅图像，这两个相关系数都在0.95以上，说明SSIM性能优越；另外，SSIM的均方根误差和背离率都小于PSNR值，对于单幅图像，SSIM的均方根误差除了一幅图像为4.3以外，其余的都小于3，说明误差很小。同时，SSIM有4幅图像背离率为0，除了一幅图像达到0.15，其余的都在0.1以内。

表2 VQEG度量指标分析

图像编号	CC		RMSE		SROCC		OR	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
a	0.842	0.961	4.701	2.412	0.810	0.946	0.238	0.095
b	0.913	0.981	5.009	2.352	0.938	0.969	0.191	0.048
c	0.885	0.976	6.348	3.000	0.915	0.962	0.091	0.046
d	0.917	0.985	5.062	2.206	0.941	0.979	0.046	0.000
e	0.893	0.977	4.919	2.334	0.949	0.970	0.095	0.048
f	0.938	0.989	3.919	1.687	0.935	0.962	0.167	0.000
g	0.858	0.979	7.286	2.913	0.949	0.971	0.191	0.048
h	0.917	0.951	5.540	4.303	0.948	0.949	0.130	0.000
i	0.907	0.950	3.968	2.945	0.937	0.950	0.100	0.150
j	0.917	0.983	4.898	2.270	0.964	0.981	0.174	0.000
全部	0.768	0.844	8.114	6.798	0.770	0.881	0.226	0.175

因此，可以认为SSIM方法与PSNR相比，更能反映人类视觉的感知效果，更适合作为静态图像压缩品质的评测标准，与现有结论完全一致，同时也验证了本文设计的评测方法结果是正确的。

同时，从图2和表1可知，SSIM在主观质量较差的时候，评估误差可能增大。因此，如何改进SSIM使其更加适合高倍压缩的品质评估是下一步工作的研究重点。

4 结束语

本文设计了一种通用的比较静态图像压缩客观评估方法的测评方法，并以SSIM方法与PSNR方法为例进行了比较，从多个角度分析，给出不同情况的度量结果，表明在主观质量相近的情况下，SSIM更加符合人眼视觉特性，与图像主观质量相对一致，适合针对静态压缩图像还原图的质量评估，与现有结论一致，同时也验证了本文设计的评测方法结果是正确的。

参 考 文 献

[1] 郑圣超. 基于HVS的若干图像质量度量方法的研究[D]. 西安：西北工业大学，2006.

[2] 杨威，赵剡，许东. 基于人眼视觉的结构相似度图像质量评价方法[J]. 北京航空航天大学学报，2008，(1)：1－4.

[3] ZHOU W，BOVIK A C，SHEIKH H R. Image quality assessment：from error visibility to structural similarity [J]. IEEE Transactions on Image Processing，2004，13 (4)：600－612.

[4] SHEIKH H R，SABIR M F，BOVIK A C. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms [J]. Image Processing，IEEE Transactions on，2006，15 (11)：3440－3451.

[5] 李永强，沈庆国，朱江，等. 数字视频质量评价方法综述[J]. 电视技术，2006，(6)：74－82.

[6] Z GUANGTAO，Z WENJUN，Y XIAOKANG. GES：a new image quality assessment metric based on energy features in Gabor transform domain [C]. IEEE International Symposium on Circuits and Systems，2006：1715－1718.

[7] RAO D V，SUDHAKAR N，BABU I R. Image Quality Assessment Complemented with Visual Regions of Interest [C]. International Conference on Computing：Theory and Applications，2007：681－687.

[8] 李波，汪海. 基于小波包变换的分层预测图像压缩算法[J]. 计算机学报，1999，(7)：685－691.

[9] VQEG. Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment [R/OL]. August 25，2003，http：//www. vqeg. org.

[10] 陈天. 基于特征提取与结构性失真的视频客观质量评估[D]. 西安：西安电子科技大学，2007.

作者简介

文义红 1977年生，2005年获大庆石油学院计算机应用专业硕士学位，现为北京航空航天大学在读博士研究生，研究方向为遥感图像压缩。

Performance Test for Image Quality Assessment

Wen Yihong Yang Kai Li Bo

(School of Computer Science and Technology，Beihang University，Beijing 100191)

Abstract There are many objective methods for quality assessment (QA) in static image compression domain，but the research on assessing QA's concidence with human quality judgements is inadequate. In order to compare their performances，an assessment method was designed. First build test image database，then obtain objective and subjective scores，finally compute all kinds of results based on the criteria. The test results prove that SSIM is more adapted than PSNR for assessing static image quality，and also verify the accurate of the proposed method.

Key words Still image Human visual system Image quality Assessment method